

DISSECTION OF GENOTYPIC AND PHENOTYPIC VARIATION IN SHRUB WILLOW
(*SALIX PURPUREA* L.)

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Fred Edward Gouker
January 2017

DISSECTION OF GENOTYPIC AND PHENOTYPIC VARIATION IN SHRUB WILLOW

(*SALIX PURPUREA* L.)

Fred Edward Gouker, Ph.D.

Cornell University 2017

Salix spp. and hybrids (shrub willow) are bred as dedicated bioenergy crops around the world, however there is still untapped potential for genetic improvement. *Salix* is a widely adapted and genetically diverse genus, but few studies have utilized this diversity for trait mapping or development of genomic tools. The studies of this dissertation focused on *S. purpurea*, a core reference species for breeding shrub willow bioenergy crops in North America, to understand the genetic basis for key traits and identify quantitative trait loci (QTL) that can be utilized for marker-assisted selection (MAS). A genetically diverse germplasm collection of 110 accessions from the Northeastern US was assembled, genotyped using genotyping-by-sequencing (GBS) and extensively phenotyped for key biomass, morphological, phenological, physiological, physical and chemical wood properties, and disease resistance across three years and three replicated experimental sites. The association population was further expanded through the addition of 157 accessions from Europe. Population genetic analysis revealed significant population stratification, subpopulation structuring and differentiation corresponding mainly to geographic regions. Phenotypic analysis of the US population showed wide variation among genotypes and revealed a majority of traits to be sexually dimorphic in favor of male plants including yield, but also a female biased sex-ratio. This suggests that the sex determination locus in *Salix* may be linked to loci responsible for growth and fitness. The natural phenotypic variation of the US population was evaluated in a genome-wide association study (GWAS),

which revealed several candidate genes for high biomass yield and traits significantly correlated to yield, as well as resistance to *Melampsora* rust. The studies conducted here advance the understanding of traits contributing to increased biomass in woody plants, lay the groundwork for validating the underlying genes, and will contribute to the development of marker-assisted selection and genomic selection to accelerate the efficiency and accuracy of breeding.

BIOGRAPHICAL SKETCH

Fred was born in Seoul, South Korea and grew up in the town of Hanover, PA where he graduated from Delone Catholic High School in 2006. He then attended The Pennsylvania State University at University Park. While at Penn State he became interested in plant molecular biology and continued to explore his interests in the plant sciences through successive summer internships with DuPont Pioneer in New Holland, PA where he became more focused on plant breeding. In 2010 he graduated with “High Distinction” with a bachelor’s degree in horticulture and a minor in biology. After graduating he was a summer intern with Syngenta Seeds in Slater, IA testing maize hybrids for herbicide tolerance. Directly after this research experience, he began graduate studies in the former Department of Horticulture (now Horticulture Section) in the field of Plant Breeding and Genetics at Cornell University. While at Cornell he attended courses and conferences domestically and internationally. He became involved in the graduate community and was the webmaster for the plant breeding and genetics graduate group Synapsis, coordinated the student garden for the Student Association of the Geneva Experiment Station (SAGES), and also served as the treasurer of SAGES. He also mentored five summer undergraduate research scholars through Cornell University’s New York State Agricultural Experiment Station Summer Research Scholars Program and also co-taught five plant science mini-courses through the Graduate Student School Outreach Program (GRASSHOPR) to kindergarten students at the West Street Elementary School in Geneva, NY. Fred will begin his post-graduate career as a postdoctoral associate as part of the *Vitis*Gen project.

ACKNOWLEDGMENTS

I would first and foremost like to thank my supervisor Dr. Larry Smart. He has been a great advisor and mentor throughout the years and has provided me with the guidance and support necessary for me to be where I am today. I also extend my appreciation to Dr. Walter De Jong and Dr. Jocelyn Rose for serving on my thesis committee and for their insightful suggestions and advice over the course of my dissertation. I owe special gratitude to Dr. Stephen DiFazio for his advice and words of encouragement throughout the years and all members of his lab for their help, especially Dr. Luke Evans for his assistance in the field. Also, thanks goes out to the entire CLEREL research staff for their dedicated assistance in field maintenance and data collection, including Dr. Terry Bates and Kelly Link. I would like to thank all of the past and present members of Dr. Smart's lab for their help and support as well as the research staff at Cornell's University Genomic Diversity Facility. I offer special thanks to Steve Gordner, Matt Christiansen, and Mark Scott for their help with field trial maintenance and harvesting and to Brian DeGasperis, Michael Rosato, Jane Petzoldt, Cody Lafler, Aaron Palmieri, Jeffrey Teague, Dawn Fishback, Lauren Carlson, Curt Carter, and Rebecca Wilk for their expert technical assistance. I would also like to thank my colleagues and fellow lab members Craig Carlson, Eric Fabio, and Michelle Serapiglia for their advice and support.

I would like to thank my friends old and new that I have made during my graduate career. I would like to thank my family for their encouragement throughout my graduate studies and for their support in the decisions I have made in my career. Most of all I thank my late father, Ronald Gouker and my mother, Judy Gouker, for the sacrifices they have made to afford me the opportunity to be where I am today.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS.....	v
LIST OF FIGURES	viii
LIST OF TABLES	xi
PREFACE.....	xiii
CHAPTER 1 - Introduction.....	1
Breeding Biomass Crops for Bioenergy	1
Importance of Breeding, Cultivating, and Commercializing Shrub Willow	2
Ecology, Population, and Genetic Structure	5
Genetic and Genomic Resources	7
Genetic Mapping Studies	8
Phenomics	12
CHAPTER 2 - Genetic Diversity and Population Structure of Native, Naturalized, and Cultivated <i>Salix purpurea</i>	14
ABSTRACT.....	14
INTRODUCTION	16
MATERIALS AND METHODS.....	20
RESULTS	24

DISCUSSION	31
CONCLUSION	37
CHAPTER 3 - Sex Ratio Bias and Sexual Dimorphism in the Dioecious Shrub Willow <i>Salix purpurea</i>	39
ABSTRACT	39
INTRODUCTION	40
MATERIALS AND METHODS	44
RESULTS	52
DISCUSSION	66
CONCLUSION	73
CHAPTER 4 - Genome-Wide-Association Study for a Suite of Bioenergy Traits in Shrub Willow (<i>Salix purpurea</i>)	74
ABSTRACT	74
INTRODUCTION	75
MATERIALS AND METHODS	78
RESULTS	82
DISCUSSION	100
CONCLUSION	105
CHAPTER 5 - Future Directions	106
REFERENCES	109

APPENDIX TO CHAPTER 2	135
APPENDIX TO CHAPTER 3	158
APPENDIX TO CHAPTER 4	193

LIST OF FIGURES

Figure 1.1 Willow breeding schema illustrating phenotypic selection, genetic mapping strategies, breeding populations, and possible molecular techniques to reduce the length of breeding and selection cycles.....	5
Figure 1.2 Benefits and limitations of association and QTL genetic mapping approaches.....	10
Figure 2.1 Geographic locations of collection sites for <i>S. purpurea</i> genotypes in this study. Samples were collected from A) wild naturalized accessions across four states in the Northeastern US and from B) wild native accessions across four countries in Eastern and Western Europe.....	21
Figure 2.2 Geographic map of collection sites for European accessions and corresponding principal component analysis (PCA) scatterplot of all individuals with positions along the first two axes with percentage of variance in brackets.....	26
Figure 2.3 Unrooted Neighbor-Joining (NJ) tree of 267 individuals based on a distance matrix derived from identity-by-state (IBS) probabilities in TASSEL from GBS data.....	28
Figure 2.4 Population stratification from STRUCTURE analysis based on consensus across 10 replications for each value of <i>K</i>	29
Figure 2.5 Scatterplot of first two linear discriminant axes showing relationship between clusters with 95% confidence ellipses that are connected by a minimum spanning tree.....	30
Figure 2.6 Heatmap and hierarchical clustering from affinity propagation (AP) analysis based on the genetic distance of <i>S. purpurea</i> genotypes.....	31
Figure 3.1 Box plots of biomass and morphological traits	57
Figure 3.2 SPAD values for monitoring nitrogen utilization in diverse <i>Salix purpurea</i> collections.	64
Figure 3.3 Least square means for leaf rust severity scores of female and male <i>S. purpurea</i>	65
Figure 4.1 Distribution map of 25,566 GBS SNP markers across 19 <i>S. purpurea</i> chromosome-scale pseudomolecules	88

Figure 4.2 A) Genome-wide association results for sex phenotypes of 251 natural accessions with 25,566 SNPs across 19 <i>S. purpurea</i> chromosomes and a 20 th naïve pseudochromosome represented by concatenated unassembled scaffolds.	95
Figure 4.3 A) Manhattan plot of associations of 25,566 SNPs for rust severity based on the SUPER model with $\lambda=1.12$	96
Figure A2.1 Diagram outlining data analysis used in this study	155
Figure A2.2 Summary of the A) total number of genotypes analyzed per population and B) allelic frequency observed across all samples for 2,287 SNPs.	156
Figure A2.3 K-means hierarchical clustering for 267 genotypes.....	157
Figure A3.1 A) Matrix of all pair-wise comparisons between traits by location within each year for Geneva, NY 2013.	174
Figure A3.1 B) Matrix of all pair-wise comparisons between traits by location within each year for Portland, NY 2013.....	176
Figure A3.1 C) Matrix of all pair-wise comparisons between traits by location within each year for Morgantown, WV 2013.....	178
Figure A3.1 D) Matrix of all pair-wise comparisons between traits by location within each year for Geneva, NY 2014.....	180
Figure A3.1 E) Matrix of all pair-wise comparisons between traits by location within each year for Portland, NY 2014.....	182
Figure A3.1 F) Matrix of all pair-wise comparisons between traits by location within each year for Morgantown, WV 2014.....	184
Figure A3.2 Matrix of all pair-wise comparisons between traits measured in 2015 for the <i>S. purpurea</i> F ₁ family.....	186
Figure A3.3 Matrix of all pair-wise comparisons between traits measured in 2015 for the <i>S. purpurea</i> F ₂ family.....	188
Figure A3.4 Correlation heatmap of traits measured in the diverse collection for the 2015 growing season.....	190

Figure A3.5 Correlation heatmap showing Pearson’s correlation coefficients (r) for all <i>S. purpurea</i> accessions (n=110).....	191
Figure A3.6 Multiple linear regression model for estimating second year post-coppice biomass yield from annual measurements	192
Figure A4.1 Genome-wide linkage disequilibrium (LD) in <i>S. purpurea</i> based on common single-nucleotide polymorphisms of minor allele frequency (MAF<0.05). Black circles correspond to average values of R^2 between physical marker locations (kb). The critical LD was defined at a distance of 1.9 kb.	193
Figure A4.2 Type 1 error plot showing the tradeoff between type I error rate (x -axis) and power (y-axis) for quantitative trait nucleotides (QTNs) with different effect sizes.....	194
Figure A4.3 Heat map of pairwise kinship among individuals included in the study.	195
Figure A4.4 Q-Q plots of $-\log_{10}(p\text{-values})$ from mixed model association analyses. Black circles correspond to MLM method, blue circles for CMLM method, and green circles for SUPER method.....	196
Figure A4.5 Manhattan plots of GWAS results for 24 traits using 25,566 common SNPs throughout the genome. The 20 th naïve pseudo-chromosome represents concatenated unassembled scaffolds. Chromosomal locations of $-\log_{10}(p\text{-values})$ for associations at each locus are shown with the green line illustrating the threshold for FDR<0.05	199

LIST OF TABLES

Table 2.1 Population genetic diversity summary statistics across SNP loci for <i>S. purpurea</i>	25
Table 2.2 Pairwise F_{ST} estimates between six natural populations and cultivar genotypes	25
Table 2.3 Analysis of molecular variance (AMOVA) between populations and samples of <i>S. purpurea</i>	27
Table 3.1 Phenotypic traits measured in the <i>S. purpurea</i> trials.....	47
Table 3.2 Mixed model results testing for genotype and locational effects on yield.....	53
Table 3.3 Means and standard deviations of phenotypic traits in the <i>S. purpurea</i> F_1 family (n=100) and F_2 family (n=482) in Geneva, NY.....	55
Table 3.4 Comparison of phenotypic traits for female and male individuals in the diverse collection of <i>S. purpurea</i> across three growing seasons.	58
Table 3.5 Comparison of phenotypic traits for male and female individuals in a F_1 <i>S. purpurea</i> family (n=100) and a F_2 <i>S. purpurea</i> family (n=482) measured in 2015 in Geneva, NY.	62
Table 3.6 Mixed model test for nitrogen utilization.....	63
Table 4.1 Broad-sense (H^2) and marker based narrow-sense (h^2) heritability estimates of phenotypic traits from the <i>S. purpurea</i> association population (n=110) at three locations.	84
Table 4.2 Statistically significant marker-trait associations for 110 genotypes and candidate genes.	90
Table 4.3 Statistically significant marker-trait associations for 251 genotypes and candidate genes for plant sex determination.	97
Table A2.1 Clone ID and source information for 267 <i>S. purpurea</i> genotypes	135
Table A2.2 Assignment of genotypes to clusters based on affinity propagation optimized grouping.	147
Table A3.1 Clone ID, sex, and source information for 110 genotypes in the diverse <i>S. purpurea</i> collection	158

Table A3.2 Experimental site characteristics for all trial locations	169
Table A3.3 Summary of phenotypic traits from the diverse <i>S. purpurea</i> collection	170
Table A3.4 Parameter estimates and significance values for multiple linear regression predictors of second year yield	173

PREFACE

This dissertation encompasses reviews and primary research of breeding and genetics of shrub willow (*Salix* spp.) including genetic diversity, sex determination, sexual dimorphism, genetic mapping, and methods of plant molecular breeding. Chapter 1 reviews the biology, breeding, and genetics of shrub willow, population genetics, and previous and current genetic mapping studies across breeding programs. Chapter 2 explores the genetic diversity of native and naturalized *S. purpurea* and examines the population structure among natural accessions from North America and Europe. Chapter 3 discusses the evolutionary biology of sex determination and sexual dimorphism in dioecious species and presents evidence for female sex ratio bias and male biased sexual dimorphism for primary and secondary traits in shrub willow. Chapter 4 covers quantitative genetics and association mapping of North American naturalized accessions of *S. purpurea* for 23 phenotypic traits using single-nucleotide polymorphism markers to identify quantitative trait loci and candidate genes for traits of interest. Finally, Chapter 5 summarizes results of the previous chapters in a holistic context and suggests future research projects for expanding the collection of natural accessions for further genetic diversity and mapping studies, suggests methods to validate markers discovered in this dissertation, and provides an outlook on future molecular breeding efforts. Readers of this dissertation will hopefully gain an appreciation for the issues surrounding the complex genetics and breeding of shrub willow and insights into broader topics in woody plants, dioecious species, and molecular breeding.

CHAPTER 1 - Introduction

Breeding Biomass Crops for Bioenergy

Over the last decade, next-generation sequencing (NGS) technologies coupled with breeding techniques have been employed to enhance commercially important traits in staple crops, such as yield, enhanced pest and disease resistance, and greater sustainability. Next generation sequencing, high-throughput genotyping, and molecular breeding methodologies, such as marker assisted selection (MAS), and genomic selection (GS) (Heffner *et al.*, 2009; Heffner *et al.*, 2010), have been applied to agronomically important crops such as rice (*Oryza sativa*) (Jena and Mackill, 2008) and maize (*Zea mays*) (Gupta *et al.*, 2009). Specifying important phenotypes depends on the priority of the breeding program and the crop species, but breeding programs are typically focused on disease resistance, increasing yield, improving nutritional quality, and/or abiotic stress tolerance. An important focus now across all agricultural fields is achieving sustainable yields, which involves maximizing output while reducing input with fewer resources and less land area (Garnett *et al.*, 2013; Wezel *et al.*, 2014). There remains considerable potential to better exploit genetic resources to produce sustainable yields. For specialty crops, such as bioenergy feedstocks, which are in their infancy in development and breeding relative to commercial staple crops, this provides a framework to rapidly harness the power of new sequencing technology and high-throughput genotyping (Poland and Rife, 2012; Yang *et al.*, 2012), and increasingly more high-throughput phenotyping (Tester and Langridge, 2010; White *et al.*, 2012) to reach these goals. Second generation lignocellulosic bioenergy crops, including woody perennials such as poplar (*Populus*) (Tuskan *et al.*, 2004; Tuskan *et al.*, 2006) and willow (*Salix*) (Smart *et al.*, 2007; Smart and Cameron, 2012), and perennial grasses like *Miscanthus* (Arnoult and Brancourt-Hulmel, 2015) and switchgrass (*Panicum virgatum*)

(Sanderson *et al.*, 1996), will be important model systems and commodity crops for research, sustainable improvement, and increased productivity.

Importance of Breeding, Cultivating, and Commercializing Shrub Willow

One of the biggest concerns facing the planet is the depletion of, and increased demand for, fossil fuels, together with the mitigation of high CO₂ levels produced from burning these non-renewable energy sources. One potential solution to this crisis is the use of renewable biomass from agricultural and forestry products. The US has recognized the need for alternative fuels and has turned to using renewable biomass as an energy feedstock to help alleviate the concerns of national security, energy independence, environmental harm, and the diminishing sources of fossil fuels. The US Department of Energy (DOE) has developed a National Biofuels Action Plan, established the Bioenergy Feedstock Development Program (BFDP) at Oak Ridge National Laboratory and brought into law the Energy Independence and Security Act (EISA) of 2007, which set a goal of reaching 36 billion gallons of renewable fuels by 2022 (National biofuels action plan, 2008). There has been an a four percent increase in consumption of woody biomass as a feedstock, and shrub willow (*Salix* spp.) was selected as one of several dedicated bioenergy crops by the US DOE to serve as a source of woody biomass for energy production (U.S. Department of Energy, 2016) .

Salix is a widely adapted and genetically diverse genus and as an energy feedstock for electricity, is carbon neutral, has a 1:55 total net energy ratio at the farm gate, improves soil characteristics, provides habitat for wildlife, and can be grown on marginal lands; all characteristics associated with being a sustainable high-yielding biomass crop (Heller *et al.*, 2003). Shrub willow is a high-yielding perennial with short rotation harvest cycles, low input

requirements, and relatively few pests and diseases (Cameron *et al.*, 2008; Smart and Cameron, 2012). Because of the broad range of environments where shrub willow can grow and the extensive geographical distribution, genotypes are likely to occur in a diversity of habitats. This geographical or environmental gradient may exert adaptive selection pressures influencing genetic and phenotypic variability that are correlated with ecological factors. This provides abundant sources of genetic diversity and functional traits to be utilized in genetic improvement for increasing yield. The pioneers for using shrub willow as a bioenergy crop are researchers in Sweden, the UK, and Denmark, where more than 20,000 ha of willow are being grown (Hanley *et al.*, 2011). For the past 20 years in the US there have been on-going research efforts to use and improve shrub willow as a biomass feedstock. Currently there are over 300 ha of willow being grown in NY alone and approval of the Biomass Crop Assistance Program (BCAP) program by the United States Department of Agriculture-Farm Service Agency (USDA-FSA) will provide farmers the opportunity to grow up to 3,500 ac of willow plantations on a commercial scale under a federally subsidized plan (Heller *et al.*, 2003; U.S. Department of Agriculture-Farm Service Agency, 2012). This will open the door to wider acceptance and economic viability of growing shrub willow for bioenergy. However, significant improvements in yield must be attained through genetic improvement. For breeding programs in NY, over 200 families of willow were produced throughout eight years with the goal of selecting for increased biomass (Smart and Cameron, 2008). The 2002 selection trial and subsequent 2005 yield trial resulted in a 40% increase in biomass compared to commercial reference clones. During the second, two-year harvest rotation the highest yielding cultivar *Salix viminalis* × *S. miyabeana*, ‘Tully Champion’, produced 23.8 oven dry tonnes (odt) ha⁻¹ yr⁻¹, a 77% increase in yield compared to the reference cultivar *S. × dasyclados*, ‘SV1’, which produced 13.4 odt ha⁻¹ yr⁻¹ (Smart *et al.*,

2005). A selection trial established in 2008, had its first three-cycle harvest in 2011, which produced 17.4 odt ha⁻¹ yr⁻¹ for the triploid hybrid (*S. koriyanagi* × *S. purpurea*) × *S. miyabeana* 05X-281-068 versus ‘SV1’ which yielded 14.0 odt ha⁻¹ yr⁻¹, a 24% yield increase as well as a 21% increase in yield compared to the tetraploid cultivar *S. miyabeana* ‘SX61’ (Serapiglia *et al.*, 2014b). Current yields obtained from a network of regional yield trials in North America have shown a 23% increase between multiple harvest rotations and 31% increases across successive years within rotation with a 14% improvement over current commercial cultivars (Volk *et al.*, 2011). Additional evidence of yield gains from one cycle of breeding has shown a 20% increase in biomass production of newly developed triploid hybrids (Fabio *et al.*, 2016). However, it is apparent that the length of time required from making controlled cross pollinations, to selection, and advancement to yield trials is at least five years (Figure 1.1). The development of marker-assisted/genomics-based selection will substantially accelerate the breeding and selection process and facilitate a better understanding of the genetic basis of biomass production, abiotic stress tolerance, disease and pest resistance, and wood density and composition, as well as identifying regulatory genes specific to controlling biomass production, yield and detecting genotype by environment interactions that contribute to trait variation (Figure 1.1)

Increases in yield will result from genetic improvement, but significant gains can only be achieved through genetic as well as agronomic improvements. Establishment costs and consistent yields are important factors effecting the adoption of shrub willow as a bioenergy feedstock and so agronomic traits contributing to improved yield must be considered in breeding efforts, including pest and disease resistance. In addition, agronomic costs must be reduced. The largest up-front cost is establishment and successful plantations require effective weed management, especially during the first two years of growth. Research trials have regularly used

post- and pre-emergent herbicides (Miller and Bender, 2010) but currently there are a limited number of Environmental Protection Agency (EPA) registered herbicides for use on shrub willow, so identifying products with minimal phytotoxicity will aid effective management of commercial willow plantations.

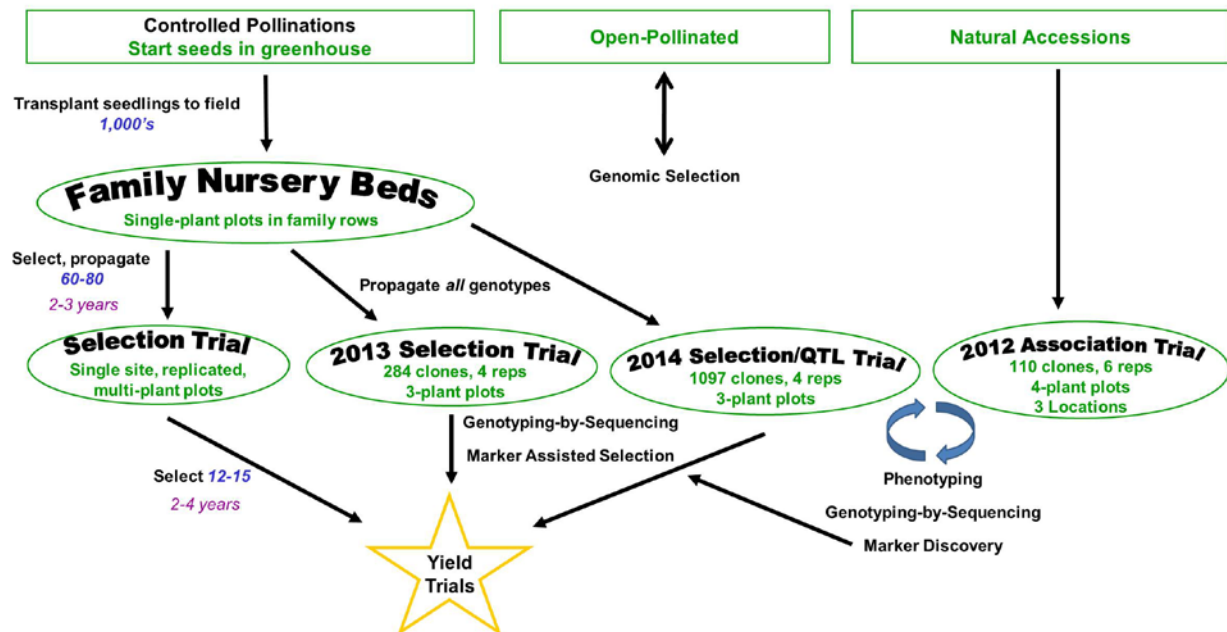


Figure 1.1 Willow breeding schema illustrating phenotypic selection, genetic mapping strategies, breeding populations, and possible molecular techniques to reduce the length of breeding and selection cycles.

In the era of genetics and genomics, there are numerous resources for implementing precision based MAS for a crop that has a long generation time and for which many traits cannot be phenotyped until several years after planting. Accelerating and increasing the accuracy of selection through molecular breeding will be essential for improved, sustainable yields.

Ecology, Population, and Genetic Structure

Natural processes, such as mutation and genetic drift, and anthropogenic disturbances of natural habitats influence the population structure and genetic diversity of species, as well as their distribution ranges. Depending on the relative importance of these processes, genetic and phenotypic differentiation can develop, where migration or introduction may lead to establishment in new habitat ranges. One way of studying processes that determine population differentiation is to use a combination of molecular and phenotypic data. This can help distinguish between stochastic and adaptive processes by testing neutral molecular markers, which can show effects of gene flow or drift. The resulting knowledge is important for predicting the ability of a species, or a population, to adapt to new environments and changing climates, or to new habitats. Knowledge of environmental adaptive potential is also important for selecting accessions and whole populations with desired traits for use in breeding programs.

These considerations are especially important in *Salix*, where there are estimated to be at least 450 different species (Argus, 2007; Lauron-Moreau *et al.*, 2015; Wu *et al.*, 2015). Many willow species have a long history of cultivation for horticultural and ornamental use, expanding the natural range of many species. Historical and traditional uses range from treatment of pain, use for cricket bats, basketry, streambank stabilization, and currently as bioenergy feedstocks. The wide utility may partially be due to the endemic nature of willow across the Northern hemisphere (Dickmann and Kuzovkina, 2008). Through introduction, primary habitat ranges have also expanded into the Southern hemisphere where only one species (*S. humboldtiana* Willdenow) exists natively in Chile and Argentina (Dickmann and Kuzovkina, 2008). Many species also thrive well in temperate areas of the South Pacific, in addition to cold climate environments in alpine and arctic regions (Jones, 1997; Beerling, 1998; Gramlich *et al.*, 2016). The broad geographical distribution of this genus has also driven ecological adaptation across all

constituent species and subpopulations within species. A high level of population differentiation has been reported across many taxa where considerable substructuring and genetic diversity is often observed (Reisch *et al.*, 2007 ; Lin *et al.*, 2009; Berlin *et al.*, 2014b; Perdereau *et al.*, 2014). Utilizing natural genetic and phenotypic variation is a powerful way to discover new traits within a species, while at the same time developing germplasm resources, and gaining insight into ecological forces driving the variation.

Genetic and Genomic Resources

Advances in genetic and genomic tools in willow have been aided by the development of resources for *Populus*, a model woody tree species for which there is a complete genome sequence (Tuskan *et al.*, 2006). The close relationship of willow and poplar as sister genera and the ongoing work in willow breeding and genomics has influenced the DOE to make significant investments, including the sequencing of the *S. purpurea* genome ("*Salix purpurea* v1.0, DOE-JGI," 2015), which is now known to have a genome size of ~390 Mb. Currently, genetic mapping studies, including F₁, F₂, open-pollinated, and association populations, are ongoing and greatly benefit from having a reference genome. Additional efforts are being made to genotype mapping populations using low-cost, high-throughput genotyping platforms (GBS) (Elshire *et al.*, 2011), to develop haplotype maps that will describe common patterns of genetic variation for use in trait mapping and marker development. The use of amplicon sequencing (AmpSeq) for genotyping and MAS will also rapidly aid in validation of single-nucleotide polymorphisms (SNP) markers for breeding programs (Yang *et al.*, 2016). The use of AmpSeq is for detecting SNPs identified from GBS and using the GBS sequence tag as an amplicon marker which can be used for subsequent genotyping. Other resources are being developed in Europe: sequencing of

the *S. viminalis* genome has been initiated in the UK by Rothamsted Research (RRes) and The Genome Analysis Centre (TGAC) (Hanley and Karp, 2013). Additional plans include resequencing of 32 more willow genomes and developing 11 QTL populations and a common garden trial for trait mapping (Hanley and Karp, 2013).

Within the last 10 years, GS has been extensively researched and developed for application in animal breeding (Dekkers, 2007; Solberg *et al.*, 2008), and has also revolutionized plant breeding strategies (Meuwissen *et al.*, 2001). This method provides fast and efficient selections where marker data across the entire genome are simultaneously used to predict the most productive genotypes. This approach has rapidly advanced the breeding of wheat (Bassi *et al.*, 2016; He *et al.*, 2016), maize (Crossa *et al.*, 2013) and soybean (Jarquín *et al.*, 2014), but GS is likely to be even more successful in long-lived perennial species, such as *Pinus* or *Eucalyptus*, where rapid artificial selection methods have not yet been applied (Denis and Bouvet, 2011; Grattapaglia and Resende, 2011; Resende *et al.*, 2012). However, for hybrid breeding programs, like those adopted for poplar and willow, which rely on heterosis, GS is not likely to work well. With the substantial genetic diversity present within *Salix*, molecular breeding and genomics-assisted selection, along with improved phenotyping platforms, will likely be superior to GS for future breeding efforts.

Genetic Mapping Studies

To date, the most widely used approach to identify genes underlying complex traits in willow has relied on bi-parental populations for QTL mapping. The concept of QTL mapping was first demonstrated by Sax (1923), based on the observation that bean (*Phaseolus vulgaris*) color, a quantitative trait, was significantly associated with the quantitative trait of bean size. However, until the development of molecular markers, there was not sufficient genetic map

coverage to dissect complex quantitative traits. With the development of a saturated molecular marker map in tomato (*Solanum lycopersicum*), it became possible to associate quantitative phenotypes with molecular markers that segregated as qualitative traits throughout the genome (Tanksley, 1988).

The first QTL maps for willow were published between 2001 and 2003 (Hanley *et al.*, 2002; Tsarouhas, 2002; Tsarouhas *et al.*, 2002; Rönnerberg-Wästljung *et al.*, 2003; Semerikov *et al.*, 2003; Tsarouhas *et al.*, 2003), but there were limitations for detecting stable QTL due to the small population sizes used. Since then, two large bi-parental families (K1 and K8), each consisting up to of 1,000 progeny, have been used for mapping loci underlying stem height, stem number, nitrogen use efficiency, rust resistance, and yield (Angela *et al.*, 2011; Shield *et al.*, 2015). In the latter case, a QTL was found on linkage group 10 (Brereton *et al.*, 2010; Hanley *et al.*, 2011). Additional studies mapped QTL for resistance to rust (*Melampsora* spp.), the primary pathogen of willow, which can reduce biomass yields by 40% (Pei *et al.*, 2008; Hanley *et al.*, 2011). A resistance locus, *SRRI*, has been identified and validated across the K1 and K8 populations, which should prove useful in identifying a range of resistance alleles across different genetic backgrounds (Rönnerberg-Wästljung *et al.*, 2008; Hanley *et al.*, 2011; Samils *et al.*, 2011; Berlin *et al.*, 2014a). Although bi-parental QTL mapping has revealed genes underlying quantitative trait variation, it is a labor intensive process that often requires development of experimental populations specific for the trait of interest. The basis of QTL mapping is to identify a statistical association between a specific genetic marker(s) and a phenotype. Bi-parental QTL mapping is effective in delimiting a gene(s) contributing to a quantitative trait to a genetic interval of 10-30 cM, depending on population size and marker density. Mapping populations are also useful for fine-mapping, and gene cloning; however, the

major limitation of QTL mapping is that it only examines two alleles of a gene and will not detect genes that have the same allele in the parents. Furthermore, it can only be applied to study the effect of the alleles in the genetic backgrounds of the mapping parents.

Genome-wide association studies (GWAS) (i.e. linkage disequilibrium (LD) mapping) is a powerful method for dissecting quantitative traits. Although this approach was initially used to map genes in humans (Hirschhorn and Daly, 2005) it has been successfully utilized to associate polymorphisms with both quantitative and qualitative traits in many plant taxa (Zhu *et al.*, 2008). Similar to QTL mapping, association mapping uses statistical models to identify markers that are linked to a trait of interest. The major differences between QTL mapping and GWAS are the choice of germplasm, the degree of LD between markers and QTL, and the number of alleles being examined (Figure 1.2).

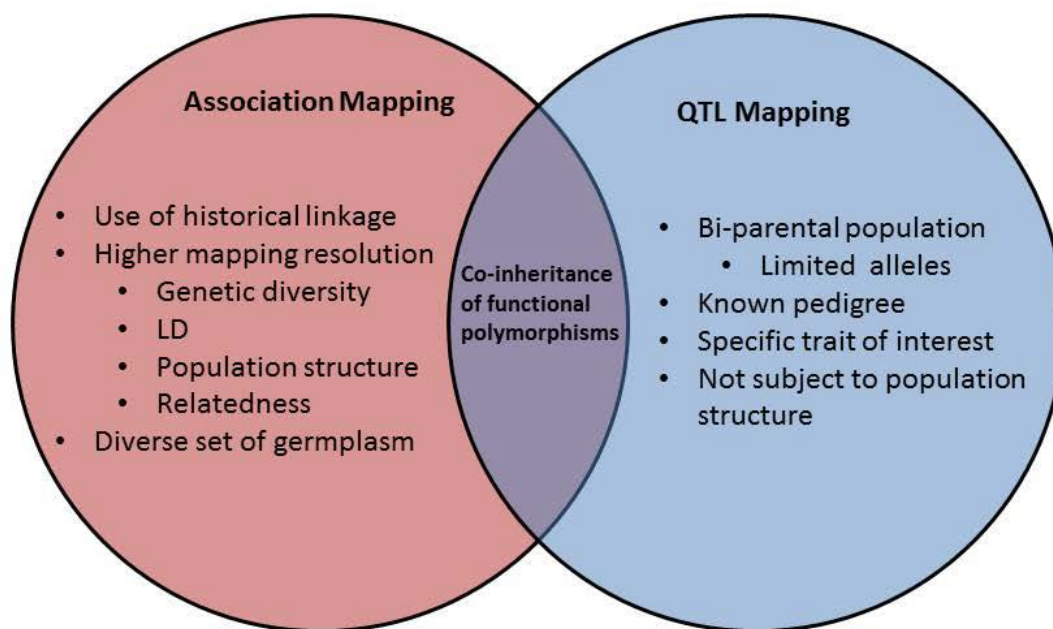


Figure 1.2 Benefits and limitations of association and QTL genetic mapping approaches.

Whereas QTL mapping is based on experimental populations derived from two parents, such that only the parental alleles are segregating in the population, GWAS utilizes natural populations, or collections of unrelated germplasm, and theoretically examines all the alleles present in that population. Both QTL and association mapping rely upon linkage of a marker to a gene controlling a trait of interest, and the resolution in both cases is limited by the number of recombination events that reduce LD between QTL and markers. Because association mapping uses natural populations, it takes advantage of all the historical recombination events that have occurred over time in the population, following divergence from a common ancestor. Depending on the germplasm used for association mapping, this approach may provide higher resolution than QTL mapping for the same number of individuals. This prediction is contingent on understanding and controlling for population substructure that may be present. In addition, association mapping can identify regions associated with a trait of interest that may not have been polymorphic in the bi-parental QTL mapping population and provides an understanding of allelic series for a trait, offering some information on the evolutionary history of the trait. Thus, association mapping is a complementary approach to QTL mapping and fine-mapping efforts in plants. The degree of LD and population structure can vary considerably between species and these differences have direct effects on association mapping. The rapid rate of LD decay in obligate outcrossing species, such as poplar, LD decays in <500 bp with ~0.5-1% nucleotide polymorphism, aids in the resolution of association mapping, but also requires dense marker coverage and a large population size (Ingvarsson, 2005). Although association mapping generally increases mapping resolution compared to QTL mapping, the resolution is limited by the extent of LD in the target region. Therefore, to identify the gene(s) underlying a QTL, fine-

mapping and experimental populations are still required. However, if association mapping is used for QTL discovery, fine-mapping can then be focused on a well resolved region.

Association mapping studies have been prominent and successful in poplar, where they utilize large populations (>500) and Infinium arrays with >30,000 SNPs or population resequencing (Slavov *et al.*, 2012; Geraldès *et al.*, 2013; Porth *et al.*, 2013; Evans *et al.*, 2014), but the genomic resources for willow are still in development. Currently, only one willow GWAS study has been published, using an *S. viminalis* population planted at two sites that used 323 accessions and 1,233 SNP markers to dissect trait variation for growth and phenology (Hallingbäck *et al.*, 2015). Further work will be needed to validate the markers and understand genotype-by-environment (GxE) interactions and phenotypic stability. The work described in this thesis has taken a similar approach, utilizing US naturalized accessions genotyped with 25,556 SNPs across three environments for three years.

Phenomics

Most traits of agricultural significance are quantitative in nature, and are controlled by multiple genes/loci. Quantitative traits are difficult to dissect genetically for several reasons: insufficient genotypic data; inadequate precision of phenotyping; and low heritability. With the technological advances in genome sequencing over the last 10 years, the limitation of dissecting quantitative traits has shifted from incomplete genetic data to insufficient phenotyping platforms, mainly referring to the efficiency and accuracy of screening large populations. The cost of genotyping has become considerably cheaper over the years and the efficiency of genotyping has become more automated, however the phenotyping in field trials is still time-consuming, labor intensive and expensive (Montes *et al.*, 2007; Singh *et al.*, 2016). To genetically dissect

quantitative traits, it is necessary to have reliable and reproducible quantitatively measured phenotypes. It is also critical that a phenomics platform is high-throughput, due to the large number of genotypes and replications necessary to obtain an accurate phenotype for each genotype (Houle *et al.*, 2010). While breeders have selected for quantitative trait improvement in breeding populations using visual evaluation and trait indexing, often based on categorical scale assessments, this approach is not sufficient to determine the underlying genetic basis of a trait. To dissect the biochemical, developmental or physiological mechanism(s) underlying quantitative traits, biologists conduct detailed experiments to evaluate fine-scaled phenotypic variation. However, it is often only possible to characterize a few genotypes at this level of resolution. Thus, more precise measurements of phenotypes are often not applicable for screening thousands of plants. To efficiently utilize the developing genomic resources, it is necessary to develop high-throughput, low cost quantitative phenotyping methods as the basis for dissecting the genetic architecture underlying traits of interest. For willow, standard protocols have been developed to accurately and efficiently measure a suite of traits broadly categorized into biomass/morphology, phenology, physiology, and physical and chemical wood properties and the protocols are presented in this thesis. Previous studies have identified stem height and total stem area to be significant predictors of overall biomass yield (Serapiglia *et al.*, 2013; Serapiglia *et al.*, 2014b; Fabio *et al.*, 2016); however, continued testing and evaluation are needed to determine if there are even more efficient measurements that predict yield reliably in a breeding program.

CHAPTER 2 - Genetic Diversity and Population Structure of Native, Naturalized, and Cultivated *Salix purpurea*¹

ABSTRACT

Salix purpurea is a woody perennial that is bred as a high yielding bioenergy crop in North America. To gain an understanding of the genotypic variation and assist with basic and applied genetic research, this study characterized the population structure and genetic diversity of *S. purpurea* from its native range of Europe and naturalized range of the Northeastern United States (US). A total of 267 genotypes of *S. purpurea* were analyzed, which included 94 naturalized accessions and 19 horticultural cultivars from the US and 154 accessions from the native range of four European countries. All individuals were evaluated using a filtered set of 2,287 genotyping-by-sequencing (GBS) single nucleotide polymorphism (SNP) markers. Using five clustering techniques (PCA, Neighbor Joining, STRUCTURE, DAPC, and Affinity Propagation) population structure results showed three broadly classified groups. Further analysis revealed seven to eight subpopulations which corresponded to geographical collection sites. As expected, the native European accessions exhibited greater diversity and subpopulation structure than the US naturalized accessions where there was a clear geographical delineation between the alpine/sub-alpine collections and the lowland collections at the Baltic Sea and Oder River. It was also shown that a subset of the horticultural cultivars was derived from the US naturalized population but also has a hybrid ancestry, likely due to introgression from breeding.

¹Chapter 2 is currently being prepared for publication and was reformatted from a manuscript with co-authors. This work was in collaboration with Steve DiFazio, Ben Bubner, Matthias Zander, Christian Ulrichs, and Larry Smart. A majority of the DNA extraction and sequencing preparation was done myself. My major contribution was carrying out the data analysis, interpretation of the results, and drafting the manuscript.

Additionally, several accessions that were thought to be distinct genotypes were found to be clonal. Ongoing and future conservation and association studies should benefit from these known substructures and diversity assessments.

INTRODUCTION

Shrub willow (*Salix* spp.) is an established high-yielding, woody perennial feedstock used in short-rotation coppicing systems in North America and Europe for bioenergy production (Smart and Cameron, 2012; Shield *et al.*, 2015). Its fast growth, perennial nature, rapid regrowth after coppice, and broad phenotypic and genotypic diversity are key traits that make it a suitable bioenergy crop. The genus is comprised of at least 450 species (Argus, 1997; Skvortsov, 1999; Wu *et al.*, 2015) and is native across the Northern and Southern hemispheres (Dickmann and Kuzovkina, 2008), where centers of biodiversity of willow are found throughout Asia, Europe, and North America (Sulima *et al.*, 2009). However, many *Salix* cultivars and natural accessions are of unknown provenance because of anthropogenic disturbance, which has also contributed to the global dissemination of genotypes (Kuzovkina *et al.*, 2008). Additional ambiguity exists because of improper identity and classification by plant taxonomists, systematists and breeders (Meikle, 1992; Rechinger, 1992; Argus *et al.*, 2010), due to small, lightweight wind-dispersed seeds, ease of natural hybridization, clonal propagation, and large inter- and intraspecific phenotypic variation (Karrenberg *et al.*, 2002; Smart and Cameron, 2012).

Many willow species have a long history of cultivation for horticultural and ornamental use, expanding the natural range of many species. An estimated 100 willow species are native to North America (Argus, 2007). Several species native to Europe and Asia that were imported to North America for ecological, forestry, and horticultural purposes have since become naturalized, including *S. alba*, *S. babylonica*, *S. fragilis*, and *S. purpurea* (Brown, 1921; Kuzovkina and Quigley, 2005). Through introduction, primary habitat ranges have also expanded into the southern hemisphere where only one species (*S. humboldtiana* Willdenow) is native in Chile and Argentina (Dickmann and Kuzovkina, 2008). Many species also thrive well

in temperate areas of Australia, New Zealand, and South Africa (Stokes, 2008), while also having adapted to harsher environments in alpine, arctic, and sub-arctic environments (Jones, 1997; Beerling, 1998). It is apparent that the broad geographical distribution of this genus is far reaching and the adaptations needed to survive these wide inhabited regions has allowed opportunities for geographic isolation and local environmental adaptation of certain species and populations.

Salix species are considered pioneers and are usually the first woody plants to colonize along lake shores, stream sides, and other low wetland areas as well as alpine regions of glacial forefronts (Hardig *et al.*, 2000; Gramlich *et al.*, 2016). The purple osier willow (*S. purpurea* L.) is a pioneer shrub that is widespread in lowland areas, with a native distribution across Europe and Northern Africa (Dickerson, 2002; Dickmann and Kuzovkina, 2008). *Salix purpurea* is prevalent both in Western and Central Europe with a wide distribution in Austria, Switzerland and Germany (Julkunen-Tiitto, 1996; Skvortsov, 1999). It is known to readily hybridize with many other *Salix* species (Argus, 1974), where crossing barriers are usually only limited by geographic isolation. Originally brought to North America by European immigrants for basketry and weaving, it has become a naturalized species commonly found throughout the Northeastern US (Brown, 1921; Dickerson, 2002). It has since become a key reference species for shrub willow breeding in the US and for which the US DOE has made significant contributions towards willow genomics, including sequencing of the *S. purpurea* genome ("*Salix purpurea* v1.0, DOE-JGI," 2015).

The study of genetic diversity of *Salix* spp. has become increasingly important not only for conservation and management, but also for serving as a germplasm resource for shrub willow breeding, where information of population structure will assist in association for improved

cultivar development (Hallingbäck *et al.*, 2015). Population diversity studies have been conducted across several species of willow for ecological and breeding interests using a variety of molecular markers including AFLPs, SSRs, and RAPDs. Many of these included shrub species, such as *S. exigua*, which was found to have a significant number of clonal genotypes within native populations on the Pacific coast of California (Douhovnikoff and Dodd, 2003; Douhovnikoff *et al.*, 2005). Significant population differentiation and high levels of genetic diversity were shown between intercontinental populations of *S. herbacea* (Reisch *et al.*, 2007; Alsos *et al.*, 2009). A great amount of genetic diversity has also been discovered between natural Irish populations of *S. caprea* (Perdereau *et al.*, 2014). *Salix viminalis* is a key species for breeding bioenergy crops in Europe and investigations of the level of genetic diversity of germplasm collections for potential use for association mapping have shown high levels of heterozygosity and geographically differentiated subpopulations across Sweden and the UK (Lascoux *et al.*, 1996; Berlin *et al.*, 2014b). To date, only four studies have been reported examining population structure and diversity within *S. purpurea*. Two studies analyzed 16 natural accessions from Poland and found greater than 70% polymorphism indicating a high degree of genetic diversity between genotypes (Sulima *et al.*, 2009; Sulima and Przyborowski Jerzy, 2013). Gramlich *et al.* (2016) examined 156 *S. purpurea* accessions from three locations in the Swiss Alps and found evidence of population structuring due to geographic isolation by river systems and mountain ranges. The study also found considerable amount of gene flow between populations of *S. purpurea* and *S. helvetica* resulting in natural hybrids. Lin *et al.* (2009) analyzed 30 subpopulations of naturalized *S. purpurea* in central NY and found high levels of genotypic diversity, but relatively low levels compared to the native species *S. eriocephala* and provided evidence that some naturalized subpopulations of *S. purpurea* have become established through

clonal propagation.

Until now, a comprehensive analysis comparing the diversity and genetic structure of native European and naturalized North American *S. purpurea* accessions has not been completed. Analysis of *S. purpurea* accessions from the native range provides an opportunity for enhanced studies of phenotype-genotype associations and insight into the genetic bottleneck that may have occurred through naturalization of *S. purpurea* in North America. Also, due to the ease of vegetative propagation, little information is known about the relationship between native and naturalized accessions and ornamental cultivars. Collections of *S. purpurea* exist in North American willow breeding programs (Zsuffa, 1990; Smart and Cameron, 2012) and throughout several European germplasm collections (Kuzovkina *et al.*, 2008). Most of these existing collections were established before the availability of molecular identification and when selection was based primarily on morphology. Molecular characterization is now widely applied across many crop germplasms and is especially valuable for those that are vegetatively propagated for which duplicate clones may exist (Urrestarazu *et al.*, 2012; Emanuelli *et al.*, 2013; Berlin *et al.*, 2014b).

This study has combined genotypes from collections of natural populations of *S. purpurea* from 41 sites in Western and Eastern Europe, 45 sites in the Northeastern US and a set of *S. purpurea* horticultural cultivars for comparison using SNP markers GBS (Elshire *et al.*, 2011). Evaluation of this collection is important for management of genetic resources, evaluating genetic diversity, and understanding population structure and relatedness of genotypes. This is also essential in the context of conventional plant breeding. Breeding methods like association mapping exploit natural variation available in natural populations. However, successful association studies require an understanding of population structure and genetic diversity. The

objectives of this study were to examine the level of genetic diversity, population structure and differentiation of the collection and identify relationships existing between native, naturalized, and hybrid cultivars with the future application of these accessions being evaluated in association mapping studies.

MATERIALS AND METHODS

Plant Sampling

A total of 267 *S. purpurea* genotypes were studied including 94 natural accessions collected within the Northeastern US (Lin *et al.*, 2009). The US population included 16 *S. purpurea* horticultural cultivars for comparison (Figure 2.1A, Table A2.1). A collection of 158 natural accessions were also obtained from the native range of *S. purpurea* spanning Austria, Germany, Poland, and Italy from 41 sites with 10-12 collections per site (Figure 2.1B, Table A2.1).

DNA Extractions and Genotyping

Young leaf tissue and shoot tips were harvested, flash frozen in liquid nitrogen, and stored at -80°C until extraction. Tissue from North American samples was collected at Cornell University, USA and was homogenized in a Geno/Grinder 2000 ball mill (SPEX SamplePrep, LLC; Metuchen, NJ, USA) and genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen; Valencia, CA, USA). Samples of European genotypes were prepared at the Thünen Institute of Forest Genetics in Germany, where frozen tissue was ground into a fine powder

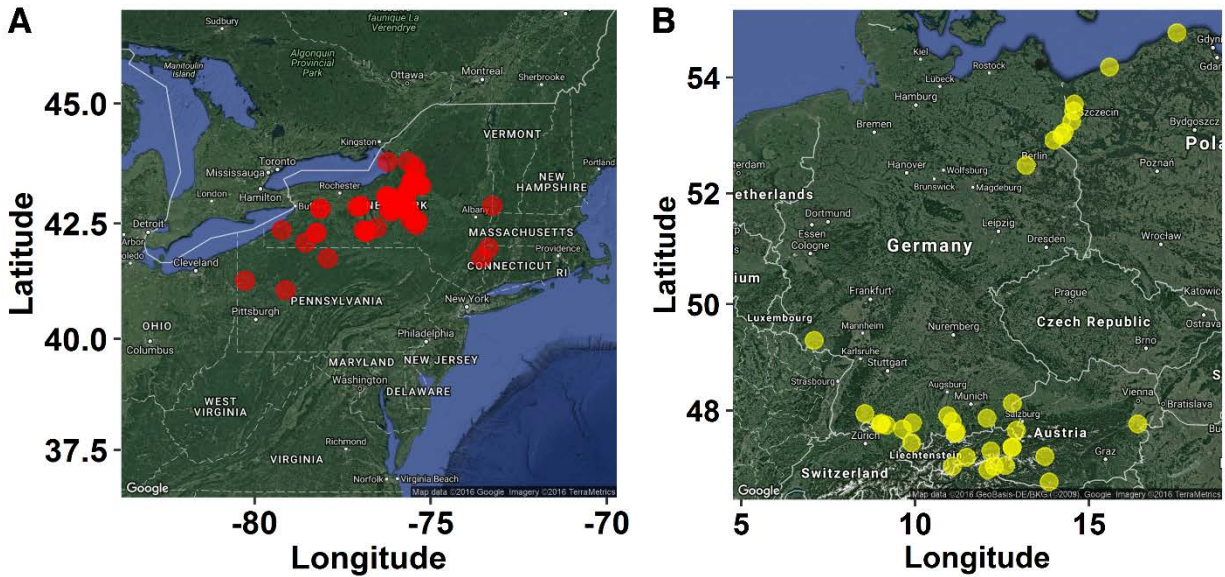


Figure 2.1 Geographic locations of collection sites for *S. purpurea* genotypes in this study. Samples were collected from A) wild naturalized accessions across four states in the Northeastern US and from B) wild native accessions across four countries in Eastern and Western Europe

with liquid nitrogen using a mortar and pestle. DNA was extracted using a modified purification protocol with dichloromethane and isopropanol (Chaves *et al.*, 1995). All DNA samples were eluted in TE buffer and final concentrations were quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific; Wilmington, DE, USA) and then normalized to 100 ng μL^{-1} . Quality assessment was performed on all DNA samples with 1 μL (100 ng) which were run on 1% w/v agarose gel along with 500 ng of λ HindIII size standards. Additional quality control was run with restriction digestion for 20 randomly selected samples using *HindIII*. Library and sequencing preparation was based on a 48-plex GBS protocol according to Elshire *et al.* (2011). DNA samples were digested with the restriction enzyme *ApeKI* due to its methylation sensitivity and uniform distribution of cut sites across the *S. purpurea* genome. The resulting libraries were sequenced on the Illumina HiSeq 2000 (Illumina, Inc.; San Diego, CA, USA) platform at the Cornell University Biotechnology Resource Center (Ithaca, NY, USA).

SNP Calling

Single nucleotide polymorphism discovery and filtering was performed with TASSEL v3.0 GBS Discovery Pipeline and a custom perl script (Bradbury *et al.*, 2007). Raw reads from FASTQ files were trimmed to 64 bp and were processed to create a set of unique sequence tags, where the minimum count that a tag must be present across all samples was set to five, which resulted in 4,550,690 unique tags. Marker genotypes were called through physical alignment to the 94006 reference genome ("*Salix purpurea* v1.0, DOE-JGI," 2015) using BWA mem module (Li and Durbin, 2009). Single nucleotide polymorphisms were retained in individuals with a call rate of <90% (removed with >10% missing data) and filtered with a minor allele frequency (MAF) <0.05, and genotypes were also screened for a minimum proportion of 50% missing data. Due to the obligate outcrossing and the highly heterozygous nature of this species, additional stringent parameters were set by removing SNPs with a call rate below 100% resulting in a final marker data set of 2,287 SNPs.

Data Analysis

Pairwise genetic differentiation among pre-defined populations was estimated with F_{ST} and was calculated with the R package *adegenet* (Jombart and Ahmed, 2011). Genetic diversity (expected heterozygosity) was calculated with the R package *mmmod* (Winter, 2012) from allele frequencies using the *glMean* function in *adegenet* (Jombart and Ahmed, 2011). An inter-individual Euclidian distance matrix was calculated from the GBS data in R and AMOVA was performed using the packages *pegas* (Paradis, 2010) and *poppr* (Kamvar *et al.*, 2014).

A dendrogram of all individuals was generated using the Neighbor-Joining (NJ) method performed with Geneious v9.1.5 (Kearse *et al.*, 2012). Clones were inferred based on visual

inspection of the dendrogram and pairwise identity greater than 84% which was determined from known clonal genotypes that were genotyped with GBS.

The genetic structure of the collection was analyzed using principal component analysis (PCA) of SNP markers implemented in TASSEL v5.2.28 and by using model-based Bayesian clustering with STRUCTURE v2.3.4 (Pritchard *et al.*, 2000a; Falush *et al.*, 2003). In STRUCTURE, the admixture model was applied and no prior population information was used. Run parameters included 10 independent replications of K values ranging from 1 to 10 with a burn-in period of 10,000 iterations followed by 100,000 Markov chain Monte Carlo replications. The results were summarized and the best estimated subpopulations were inferred based on plotting the log probability $L(K)$ and ΔK according Evanno *et al.* (2005) over ten runs using Structure Harvester v0.6.94 (Earl and vonHoldt, 2012). Independent runs were aligned using CLUMPAK (Kopelman *et al.*, 2015) and clustering results were visualized using DISTRUCT v1.1 (Rosenberg, 2004).

A multivariate method of clustering was also evaluated by discriminant analysis of principal components (DAPC) using the R package *adegenet* (Jombart *et al.*, 2010) with the functions *glPca*, *find.clusters*, and *dapc* functions. The *n.start* argument and *find.clusters* using the sequential k -means clustering algorithm was used to make the function converge on a single answer for six or seven clusters, respectively. The first 100 principal components were retained for DAPC analysis based on the recommendation in the *adegenet* documentation. Bayesian information criterion (BIC) was used to infer the optimal number of genetic clusters minimizing variation within and maximizing variation between identified groups.

The affinity propagation (AP) clustering method (Frey and Dueck, 2007) was also used to assess genetic clusters and compare with those identified by STRUCTURE and DAPC methods.

Analysis was conducted in R using the *apcluster* package (Bodenhofer *et al.*, 2011) where 100 independent runs were conducted to validate the identified *exemplar* centers for each cluster. This approach allowed assessment of the robustness of previously identified subpopulations and identify additional subpopulations or provide reassignment of the accessions to new or existing clusters. An overview of the data analysis workflow is provided (Figure A2.1).

RESULTS

Genetic Diversity and Differentiation

Population genetic diversity analyses revealed noticeable differences between the US and cultivar genotypes and the European natural accessions (Table 2.1, Figure A2.2). Across sampling regions, the overall number of different alleles was significantly greater in the European accessions than the US accessions and cultivars ($P < 0.05$), whereas there was no difference in the allelic richness between the US accessions and cultivars. Mean heterozygosity observed across all markers was 0.248 and mean expected heterozygosity was 0.250. There were no significant differences between the observed and expected heterozygosity, but there was a significant deviation from Hardy-Weinberg equilibrium for the majority of markers ($P < 0.05$), which can be seen in the heterozygote excess, especially with the negative F_{IS} values observed within the US accessions and cultivars.

Table 2.1 Population genetic diversity summary statistics across SNP loci for *S. purpurea*^a

Sampling Region	N	N_a	H_o	H_e	F_{IS}
US	94	1.47	0.276	0.257	-0.034
Europe-Alpine	100	1.83	0.221	0.266	0.073
Europe-Baltic	54	1.85	0.234	0.229	0.011
Cultivar	19	1.46	0.260	0.247	-0.021
Sampling Population					
U.S	94	1.47	0.276	0.257	-0.034
Alpine Foothill	67	1.41	0.237	0.227	0.011
Alpine	33	1.37	0.207	0.207	0.032
Baltic Sea Coast	2	1.30	0.172	0.213	0.039
Baltic-Oder River	42	1.42	0.231	0.230	0.016
Baltic Inland	10	1.41	0.247	0.229	-0.052
Cultivar	19	1.46	0.260	0.247	-0.021

^aN is the number of individuals sampled, N_a is the mean number of alleles, H_o is the average observed heterozygosity, H_e is the average estimate of heterozygosity, F_{IS} is the fixation index.

When populations were considered within sampling regions, allelic richness was reduced in the European accessions when divided into the five separate populations. All fixation indices remained positive in the European accessions except for those from the Baltic Inland population that had a negative F_{IS} value (-0.052, Table 2.1). Significant genetic differentiation between all populations was observed as indicated by the F_{ST} values (Table 2.2) with a mean overall F_{ST} value of 0.160.

Table 2.2 Pairwise F_{ST} estimates between six natural populations and cultivar genotypes

	US	Alpine Foothill	Alpine	Baltic-Sea Coast	Baltic-Oder River	Baltic - Inland
Alpine Foothill	0.210					
Alpine	0.258	0.103				
Baltic Sea Coast	0.385	0.422	0.417			
Baltic-Oder River	0.159	0.184	0.242	0.369		
Baltic Inland	0.159	0.187	0.239	0.378	0.065	
Cultivar	0.102	0.142	0.203	0.406	0.174	0.171

Additionally, PCA revealed clustering of the three sampled regions and further clustering and separation of the six natural populations corresponding to geographic location with hybrid

cultivars scattered amongst the natural accessions (Figure 2.2). Population differences were apparent in the PCA where dense clusters were observed with Alpine and Baltic accessions, whereas the US accessions were scattered over a wider range, but also overlapped with the European accessions along the PC2 axis. The positioning of two F_1 hybrid cultivars ('Fish Creek' and 'Wolcott') between US parental accessions (94006 and 94001) provides genetic validation of the marker dataset used for clustering analysis. The genetic differentiation between the US accessions and cultivars was the second lowest ($F_{ST}=0.102$) compared with the lowest genetic differentiation seen between the Baltic-Oder River and Baltic Inland populations ($F_{ST}=0.065$). The greatest genetic differentiation was between accessions from the Alpine

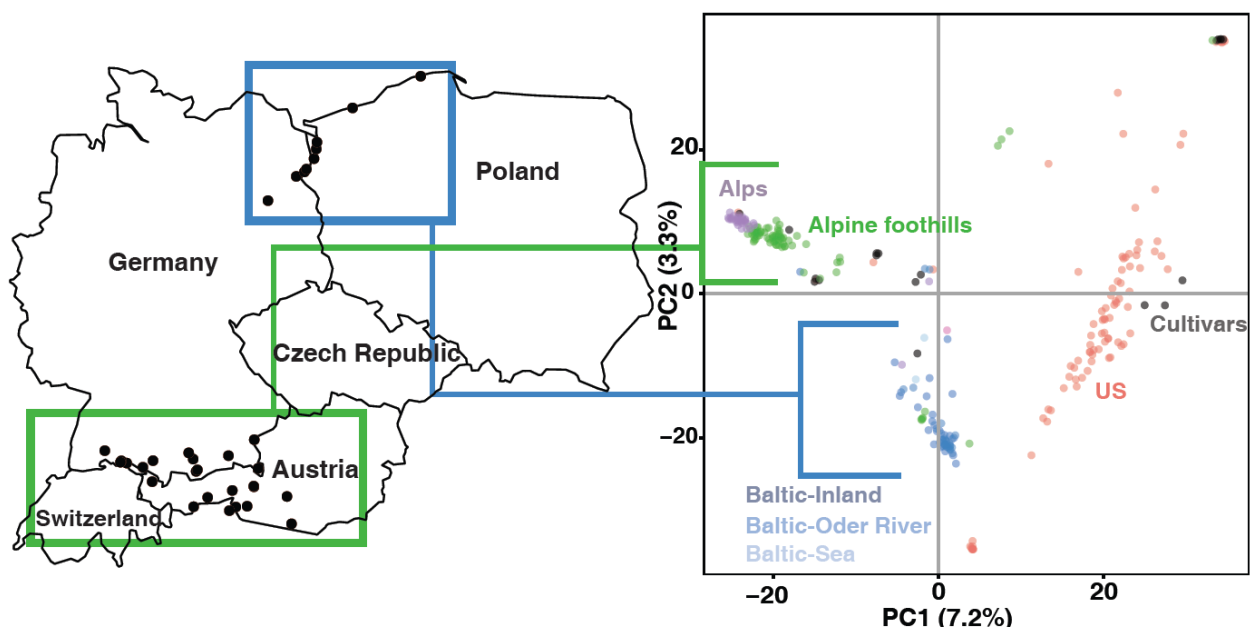


Figure 2.2 Geographic map of collection sites for European accessions and corresponding principal component analysis (PCA) scatterplot of all individuals with positions along the first two axes with percentage of variance in brackets.

foothills and the Baltic Sea Coast with a similar level of differentiation between the Baltic Sea Coast the Alpine derived accessions. There were also similar levels of genetic differentiation between the Baltic Sea Coast accessions and the other two Baltic populations as there was

between the Baltic Sea Coast and the US, Alpine, and Alpine foothill accessions. The greatest population differentiation between the cultivars and a natural population was with the Baltic Sea Coast. The US population compared with other the natural populations, also showed the greatest level of differentiation with the Baltic Sea Coast. A global analysis of molecular variance (AMOVA) showed a higher level of variation among populations (9%) than between samples among populations, which is reflected in the dense clustering of the European accessions. However, most of the variation explained (87%) came from between samples within the populations.

Table 2.3 Analysis of molecular variance (AMOVA) between populations and samples of *S. purpurea*^a

Source of Variation	df	Sum of Squares	Mean Squares	Estimated Variance	Explained Variance (%)
Among Populations	3	10832.831	3610.944	29.470	9
Between Samples among Populations	263	68468.787	260.338	12.906	4
Between Samples within Populations	267	75735.500	283.654	280.707	87
Total	533	155037.118	290.876	322.652	100

^adf, degrees of freedom

Genetic Relationships and Population Structure

To infer genetic relationships of individual genotypes and identify clones or highly related individuals, an unrooted NJ tree was constructed (Figure 2.3). The NJ tree revealed four major groups identified by main geographic regions with distinct branching between the US, Baltic, Alpine foothill, and Alpine accessions. There were no distinctive outgroups identified. There was not a clear separation of nodes between the Baltic accessions, where there was a greater degree of node branching among the US and Alpine accessions with the greatest branch length occurring between these two groups. There were five pairs of known clonal genotypes that were sequenced with GBS, and pairwise percent identity between the clonal pairs (>84%)

was used to infer other clonal genotypes across all samples, which revealed 6% of the genotypes to be clonal or have a high degree of genetic similarity. A pair of full-sib cultivars used in the analysis showed 80% pairwise identity to each other and 79% pairwise identity to the parental genotypes. The greatest degree of genetic similarity (94%) was seen between four accessions from the Alpine region (PU49, 50, 51, 52).

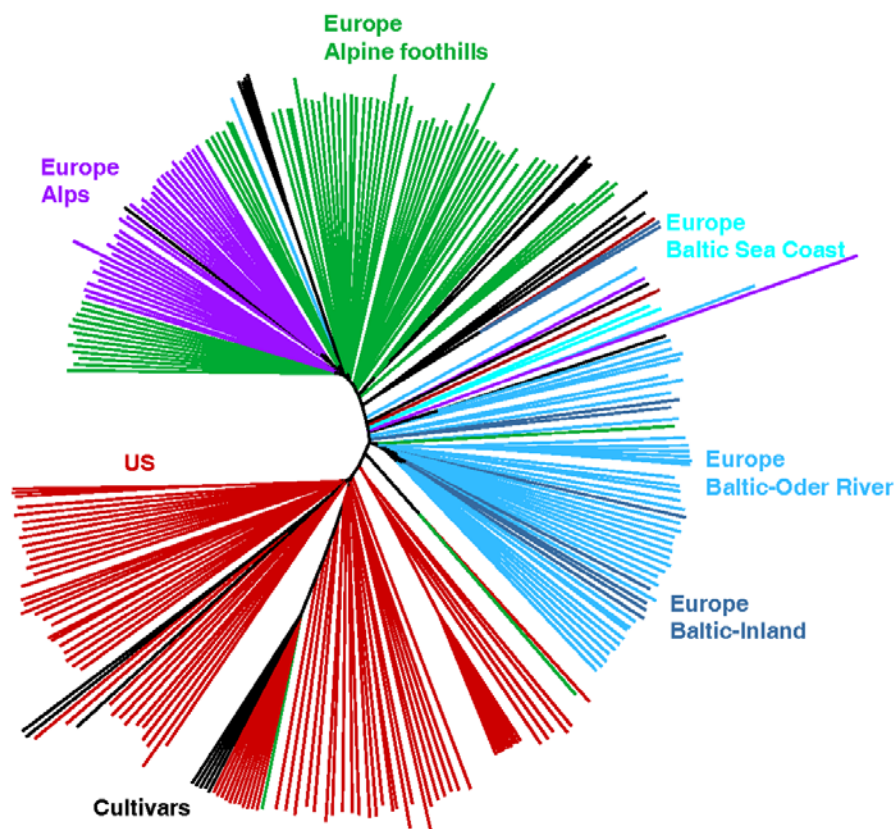


Figure 2.3 Unrooted Neighbor-Joining (NJ) tree of 267 individuals based on a distance matrix derived from identity-by-state (IBS) probabilities in TASSEL from GBS data

The Bayesian clustering performed with STRUCTURE supported the optimal number of genetic clusters to be between two and eight ($K=4$ to 8) (Figure 2.4). Based on log-likelihood values the optimal value was $K=8$, while the optimal number of clusters according to ΔK was 2. Distinct groupings were seen based on broad geographic regions and proportion of membership

of individuals increased across clusters with increasing values of K , with the greatest evidence of admixture seen in the European-Baltic accessions.

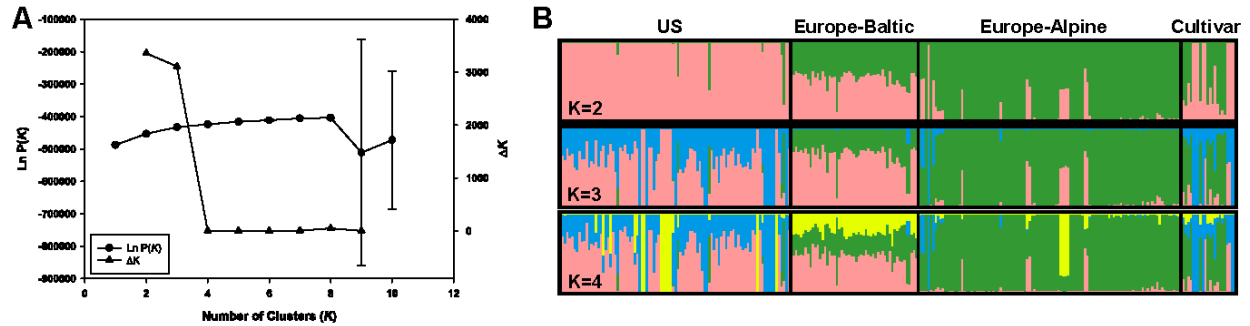


Figure 2.4 Population stratification from STRUCTURE analysis based on consensus across 10 replications for each value of K . A) Circles with standard deviations show average log-likelihoods across independent runs for each K . Triangles indicate values of Evanno's ΔK based on the rate of change of the log-likelihood. B) Bar plots representing population structure. The number of clusters is shown for $K=2$ to $K=4$. Vertical bars represent each genotype and length of the colored bar shows the estimated proportion of membership to each cluster.

Additionally, DAPC analysis revealed an optimal number of genetic clusters of seven ($K=7$) (Figure 2.5) based on the BIC values obtained from sequential K -means clustering (Figure A2.3B). Retaining 100 PCs and using the first two linear discriminants accounted for 79% of the total variance. Assignment of individuals to each group was congruent with that of the STRUCTURE analysis except that the US accessions were separated into two clusters suggesting subpopulation structure. However, there still a high level of genetic similarity between these clusters based on short edges of the minimum spanning tree at the center of the clusters. A similar trend was seen with the Alpine foothill and Alpine accessions, but with a greater overlap of individuals between clusters. The Baltic Inland and Oder River accessions were grouped together in a DAPC cluster and were at the center of the minimum spanning tree between clusters similar to what was seen in the NJ tree.

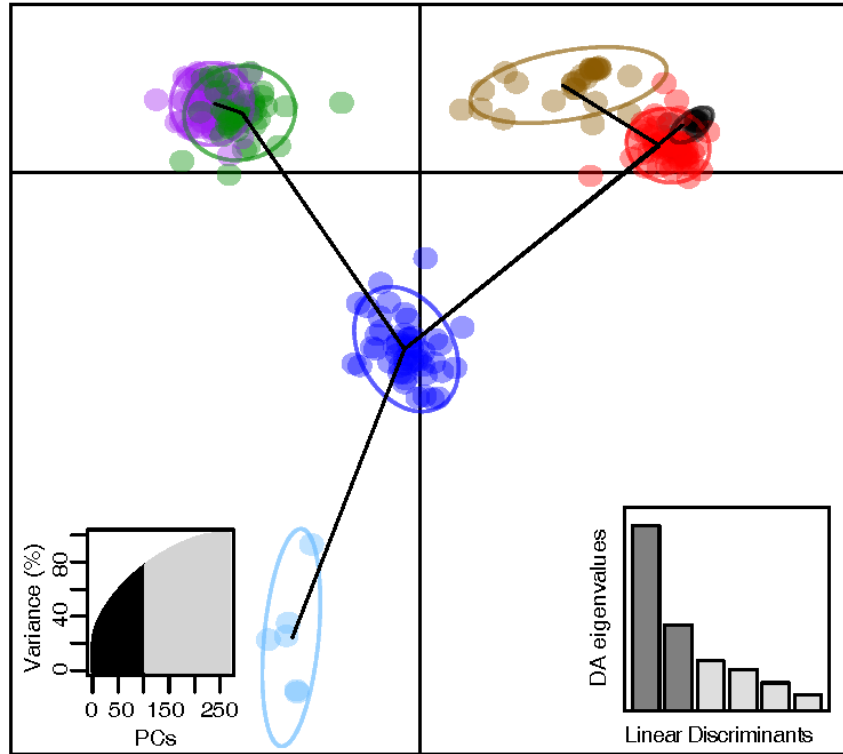


Figure 2.5 Scatterplot of first two linear discriminant axes showing relationship between clusters with 95% confidence ellipses that are connected by a minimum spanning tree. Eigen values are shown for the first 200 principle components (PCs). Dark blue circles indicate clustering of predominantly Baltic Inland and Oder River populations, light blue for the Baltic Sea population, green for the Alpine population, purple for Alpine foothill, brown and red for US subpopulations, and black for cultivars

Lastly, an AP analysis was performed to examine further refinement of genetic clusters. Unlike STRUCTURE and DAPC, AP does not require prior estimates of K . Based on hierarchical clustering (Figure 2.6) there was consistent genetic homogeneity in groups one, two, six, and seven consisting mainly of US accessions and cultivars (Table A2.2) with a portion of the remaining US accessions assigned to the other groups. Cluster three consisted of a mixture of cultivars, Alpine foothill, Baltic Inland accessions, and one US accession (03-01-036) serving as the exemplar for that group. Group four assignments were mainly Alpine and Alpine foothill accessions and group five consisted mainly of Baltic Inland and Baltic Oder River accessions. The heatmap was based on negative Euclidean distances, and grouped accessions based on

genetic distance profiles to all other accessions (Figure 2.6). The order of the arranged clusters on the heatmap were joined agglomeratively based on exemplars. Although the AP results suggested seven clusters, the heatmap indicated six clusters based on groupings along the diagonal where groups six and seven are encompassed by the same light colored block.

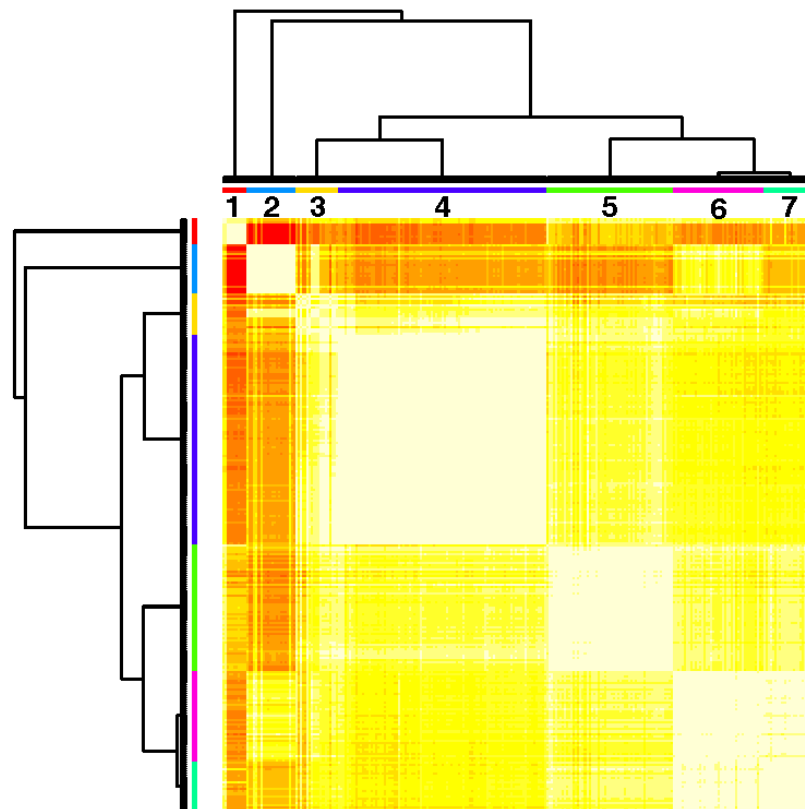


Figure 2.6 Heatmap and hierarchical clustering from affinity propagation (AP) analysis based on the genetic distance of *S. purpurea* genotypes. Yellow color indicates low degree of similarity between genotypes and orange indicates a high degree of similarity

DISCUSSION

The dispersion of naturalized *S. purpurea* in North America is unknown due to the historical introduction from regions of its native range found throughout many European countries. Because of its importance as a biomass crop for bioenergy production in North

America, it has been widely used in breeding programs and more recently in genetic mapping studies (Smart and Cameron, 2012; Serapiglia *et al.*, 2013). In order to develop superior cultivars with improved yield, mapping studies such as association analyses will need to be implemented. However, the level of genetic diversity and population structure are confounding factors in association genetics and need to be well characterized beforehand. In this study, I report on the genetic diversity and population structure of native and naturalized accessions as well as comparing cultivars to seek inference on hybrid origins using a set of 2,287 GBS SNPs. The result from this study indicate significant population structuring as well as identification of several clonal genotypes.

Clonal identification was expected within *Salix* populations due to the ease of vegetative reproduction of shrub willow and the close proximity for some of the sampling locations. Seven pairs of accessions were found to be clonal with all but one pair coming from the European population. The 84% pairwise identity threshold for clonal genotypes is lower than what is expected. However this is due to the level of missing data generated by the GBS method, and therefore the level of similarity for clones is a technical limitation of GBS. The number of clonal genotypes identified was proportionately lower than other studies of willow using natural populations. Sixty-seven of the US accessions examined in this study were previously characterized using nine microsatellite markers where five accessions were identified as clones (Lin *et al.*, 2009). Although several accessions in this study had high levels of pairwise identity and clustered similarly on the NJ tree, the level of similarity was not as great as known clonal genotypes or F₁ full-sibs and therefore were not considered as clones. However, relatedness and the close genetic similarity may be due to several generations of interbreeding among related individuals. The greater than 94% identity between European Alpine accessions (PU49, PU50,

PU51, PU52) indicated clonal genotypes. This would be expected with accessions collected from the same sampling locations, as in this case. Other genotypes identified as clones were also found to be collected from separate sites (PU118 and PU120) however, the close proximity of 100 km between sampling locations along the Kalserbach Creek suggests vegetative propagation is likely to be the reason. It is known that high levels of vegetative propagation occur naturally in *Salix* due to rooting of stems that have traveled downstream in river beds (Moggridge and Gurnell, 2009; González *et al.*, 2010). However, the lack of clonality among the accessions suggests establishment of the populations mainly occurred through sexual reproduction and seed dispersal. The proportion of sexual reproduction versus vegetative propagation in *Salix* depends on the proximity of females and males, the presence of pollinators, and proper environmental conditions required for seed germination (Karrenberg *et al.*, 2002). Additionally, the high degree of genetic similarity of several accessions within the same population may also be due to a combination of sexual and asexual genotypes (Shafroth *et al.*, 1994). The cultivar genotypes examined in this study were found to be similar to those of natural accessions across all populations. Many of the cultivars had greater genetic similarity with the US and Alpine accessions. This suggests that these cultivars were likely natural accessions from those particular regions and propagated for ornamental use based on unique morphological characteristics. Similarity between the cultivars and the US accessions suggests that the naturalized populations were derived from the cultivars; however the cultivars were developed after they were already naturalized, therefore it is still most parsimonious to assume that they share a common ancestral cultivar. They also showed a high degree of similarity to each other which corresponded to the origin of the commercial nursery where they were obtained, suggesting that some may be siblings developed through breeding efforts. As expected, three cultivars with known pedigrees

(‘Fish Creek’, ‘Wolcott’, 05X-293-047) clustered closely to the parental naturalized US accessions (94006, 94001, 05-01-002) which provides validation of the marker dataset used for the analysis.

Clustering analysis revealed distinct population structure corresponding mainly with geographic regions. Similar results of distinct population structure are commonly seen across the genus and has been previously described for *S. purpurea* and *S. eriocephala* (Lin *et al.*, 2009) and several other *Salix* spp. (Lascoux *et al.*, 1996; Beismann *et al.*, 1997; Reisch *et al.*, 2007; Trybush *et al.*, 2012; Berlin *et al.*, 2014b). Prior assumptions of populations were based on the three broad geographic regions where the accessions were collected. Bayesian clustering based on STRUCTURE analysis confirmed this grouping, however most of the admixed individuals grouped in the European-Baltic accessions. The hypothesis was that the hybrid cultivars would be the most admixed, but the unexpected STRUCTURE results might be due to violation of too many of the underlying assumptions used with this method. Clustering based on PCA analysis revealed refinement of the populations based on distinct collection sites from geographic regions where European Alpine accessions clustered into Alpine and Alpine foothill subpopulations. A lesser distinction was seen with the Baltic accessions. The NJ tree and DAPC analysis showed similar results with these clusters based on population sampling sites. The result of DAPC and PCA were also similar for inferring genetic structure. DAPC, however, is a semi-model based method (Jombart *et al.*, 2010) that first performs PCA, followed by *K*-means clustering based on informative PCs, and then performs a discriminant analysis to infer membership probability to individuals for each cluster which provides a robust estimate of subpopulation structure. The DAPC results still identified Alpine and Alpine foothill groups in addition to distinct subpopulations of Baltic accessions corresponding to those found in the Baltic Inland region and

along the Oder River. Similarly, US accessions were split into subpopulations corresponding mainly to separate river systems and watersheds found in central NY as previously identified (Lin *et al.*, 2009). An alternative method to identify population substructure was through the use of AP analysis. The AP results showed the same number of clusters as the DAPC and NJ methods, but identified further substructure within the US accessions, which can be partly explained by similar origins of accessions along stream banks. Affinity propagation analysis is a fairly new clustering method (Frey and Dueck, 2007) that has yet to be widely adopted in the plant population and genetics community, but which has been demonstrated to be useful in gene expression studies (Leone *et al.*, 2007; Kiddle *et al.*, 2010).

The small number of studies that have attempted to characterize the population genetic structure of various willow species, including *S. purpurea* (Lin *et al.*, 2009; Gramlich *et al.*, 2016), have relied primarily on AFLP and SSR markers. However, use of a small number of neutral molecular markers in these studies results in very limited subset of the genome, limiting the resolution and clear assessment of structuring (DeFaveri *et al.*, 2013). Single nucleotide polymorphisms are easily identified, are co-dominant, and occur in such high numbers that in some cases they can be more efficient than SSR, because they do not require large sample sizes to characterize variation within a population. The more closely related a group of individuals are, the more restriction site loci they will have in common throughout their genome. Because of this principal, GBS has been demonstrated as a valid tool for genomic diversity and population genetic analysis (Morris *et al.*, 2011; Lu *et al.*, 2013; Peterson *et al.*, 2014). The ability of this method to identify a large number of SNP loci in a genome has the possibility to overcome the limitations of traditional markers. Despite this promising potential, GBS has not been widely applied to analyze population genetic diversity across the greater than 450 species of *Salix*, but

population genetic parameters of this study still correspond with previous evaluations, thereby validating their use for this purpose.

Negative F_{IS} values were observed for US, Baltic Inland accessions, and cultivars which indicate an excess of heterozygotes, while positive values were observed for all other populations. This suggests that naturalized accessions in the US population are less homozygous than expected with random mating. This may be a consequence of the introduction of divergent cultivars from different parts of the native range, which became inbred and then recently hybridized during the naturalization process. Displacement and introduction of naturalized willow populations in the US may have come from many different source populations with the initial spread due to early long distance anthropogenic dispersal used for basketry and streambank stabilization (Brown, 1921). Conversely, the positive F_{IS} values observed in the European populations may be due to geographic isolation impeding gene flow between populations resulting in continued interbreeding amongst individuals in isolated regions. Heterozygote deficiency is also a common feature of GBS markers, so this could also be a contributing factor (Glaubitz *et al.*, 2014).

The overall F_{ST} value indicated statistically significant population differentiation and is consistent with reports for *S. purpurea* (Lin *et al.*, 2009; Gramlich *et al.*, 2016). This trend has also been commonly observed in several other *Salix* spp. showing differentiation between most subpopulations (Reisch *et al.*, 2007; Trybush *et al.*, 2012; Berlin *et al.*, 2014b; Nagamitsu *et al.*, 2014). The Baltic Sea Coast population was the most differentiated, however it had the lowest observed heterozygosity and mean number of alleles, but this was likely because of the small sample size. The lowest pairwise F_{ST} value was between the Baltic Oder River and Baltic Inland population that close in geographic proximity. However, the level of differentiation between them

suggests impeded gene flow due to pollen competition from local male plants, most likely because of stable populations along the Oder River streambank where growing conditions for establishment are optimal (Slavov *et al.*, 2009; Setsuko *et al.*, 2013). Previous estimates of F_{ST} values between US accessions of *S. purpurea* subpopulations were significantly different from zero and AMOVA analysis indicated 84% of total genetic variance occurred within subpopulations (Lin *et al.*, 2009). These results are consistent with those observed in this study with US population F_{ST} estimates ranging from 0.159 to 0.385 between populations and with 87% of the total genetic variance being explained from within each population showing a high level of genetic differentiation in the naturalized subpopulations. F_{ST} estimates between the US population and the Alpine populations were greater than the Baltic populations. This high level of differentiation was also reflected in the NJ tree and DAPC clustering with the greatest genetic distance occurring between US and Alpine populations and the Baltic population being intermediary. This suggests that the naturalized accessions in the US have a greater probability of being derived from the European Baltic region and since the pairwise F_{ST} was significantly different from zero, it indicates that the naturalized populations have greatly differentiated since the time of introduction to the US. It is also possible that sampling in the European native range did not include all of the source populations representing the origins of North American introductions and therefore the true genetic origin has not yet been captured.

CONCLUSION

This was the first population genomics study of *S. purpurea* that considered accessions from the native and naturalized range. Additionally, the greater marker density obtained from GBS (Elshire *et al.*, 2011) proved to be a reliable marker system to estimate genetic variation and

population structure where the genotypes collected in this study will serve as an association mapping population in future studies. There was evidence of population structure which can cause spurious phenotype-genotype associations and must be accounted for in any association genetic analysis (Ioannidis *et al.*, 2009). Even low levels of population structure ($F_{ST} \leq 0.01$) can increase false positive associations and can occur unless an adequate number of markers are used to control for this effect (Price *et al.*, 2006). These analyses revealed the presence of closely-related and even clonal individuals that must be removed from association studies, thus decreasing the population size and the power of the analysis. Based on these results, efforts should be made to expand the *S. purpurea* association population by collecting from a larger area in the native and naturalized ranges.

CHAPTER 3 - Sex Ratio Bias and Sexual Dimorphism in the Dioecious Shrub Willow *Salix purpurea*¹

ABSTRACT

Quantification of sex-related phenotypic differences in dioecious plants can provide evidence for the evolutionary, ecological and molecular basis of sex differentiation. In *Salix* spp., female-biased sex ratios are often observed, but the mechanisms are not well understood. Evolutionary and ecological theory suggests that males and females of a dioecious species will become sexually dimorphic leading to differences in growth, reproductive success, and fitness. This divergence should then lead to measurable differences in phenotypes, which can skew sex ratios due to differential performance of males and females. This study examined a suite of morphological, phenological, physiological and wood composition traits of *S. purpurea* L. in F₁ and F₂ families produced through breeding and in a collection of unrelated genotypes to test for sexual dimorphism and sex ratio bias in a species that uses a ZW system of sex determination. Results showed significantly greater means for a majority of traits in male genotypes as evidence for sexual dimorphism in *S. purpurea*, where many of the morphological phenotypes measured across multiple years were highly predictive of biomass yield. Yield was significantly greater in male plants, which also exhibited greater nitrogen utilization under fertilizer amendment, but also greater susceptibility to fungal infection by *Melampsora* spp.

There were also significant female-biased sex ratios in both F₁ and F₂ families. These results

¹Chapter 3 has been formatted as a manuscript with co-authors, which is being prepared for submission: Gouker, F.E, Carlson, C.H., Evans, L.M., Smart, C.D., DiFazio, S.P., Smart, L.B. (2016). Sex ratio bias and sexual dimorphism in the dioecious shrub willow *Salix purpurea*. My contribution to this work was performing the majority of the data analysis and writing the manuscript.

provide evidence for female-biased sex ratios in *S. purpurea*, but also for sexual dimorphism in favor of males expressed as greater overall growth and nutrient uptake. These data provide a foundation for further examination of the evolution and selection for a ZW sex determination system in the Salicaceae.

INTRODUCTION

Phenomena such as dichogamy and dioecy have evolved in plants to prevent self-pollination, encourage cross-pollination, and stimulate genetic diversity. However, dioecy is relatively uncommon in plants (Renner, 2014; Charlesworth, 2015) compared with monoecy or hermaphroditism. It is estimated that approximately 14,600 species in 200 families are dioecious (Ming *et al.*, 2007; Yang *et al.*, 2015), whereas the rare instance of subdioecy has been reported to occur in 32 species in 21 families (Ehlers and Bataillon, 2007). It is theorized that dioecy in flowering plants evolved from an ancestral co-sexual state prior to the development of distinct sex chromosomes (Charlesworth, 2013), a hypothesis that is supported by the prevalence of hermaphrodites among a majority of all flowering plant species (Ming *et al.*, 2007). The first account of flowering plants in the fossil records appeared approximately 124.6 million years ago (MYA) (Sun *et al.*, 2002). Even though it is likely ancestral flowering plants were hermaphrodites, it is known that dioecy has evolved independently in several plant families (Ming *et al.*, 2007) and even at the genus and species level (Westergaard, 1958). In conjunction with the evolution of dioecy is suppressed recombination and selection for sex chromosome dimorphism (Ming *et al.*, 2007; Chen *et al.*, 2016) leading to the development of homomorphic and heteromorphic sex chromosomes. The latter is less common in plants than in animals. Only a few heteromorphic chromosomes have been observed in plants with the best characterized

occurring in *Silene latifolia* (Charlesworth, 2002, 2013), which is thought to have evolved only relatively recently, <20 MYA (Bergero *et al.*, 2008). In light of the limited evidence for heteromorphic sex chromosomes, it is likely that homomorphic sex chromosomes are more prevalent in dioecious taxa.

Recent studies have identified sex-determination regions (SDRs) in dioecious species including grape, papaya, and persimmon (Fechter *et al.*, 2012; Wang *et al.*, 2012; Akagi *et al.*, 2014). These species use an XY sex determination system, characterized by male heterogamety, similar to the system used in a vast majority of mammals (Cortez *et al.*, 2014). Several dioecious species instead exhibit a ZW sex determination system (Wolf *et al.*, 2001; Spigler *et al.*, 2008), where females are the heterogametic sex, similar to what is observed in many avian taxa (Zhou *et al.*, 2014). Unlike animal systems, the exact mechanism of sex determination of most dioecious plants still remains largely unknown. However, in the past 20 years increased interest in the mechanisms of sex determination among dioecious plants has pointed to possible genetic, epigenetic, environmental, and hormonal control of sex development (Ming *et al.*, 2011; Renner, 2014).

Species in the family Salicaceae are part of the estimated 4-10% of all flowering plants that exhibit complete dioecy (Ainsworth, 2000; Charlesworth, 2002; Dickmann and Kuzovkina, 2008; Renner, 2014; Charlesworth, 2015). This includes *Populus* and *Salix*, which are dioecious woody perennials where male and female plants are discernable through clear morphological differences between staminate and pistillate catkins (Dickmann and Kuzovkina, 2008). The evolution of dioecy in the Salicaceae is thought to have occurred prior to the divergence of *Salix* and *Populus* approximately 65 MYA (Collinson, 1992; Tuskan *et al.*, 2006; Filatov, 2015). The Salicaceae appear to have indistinguishable homomorphic chromosomes based on multiple

cytological studies of *Populus* (Peto, 1938; van Buijtenen and Einspahr, 1959). In *Populus*, mapping studies indicate that sex is determined either using an XY or ZW system depending on the species (Pakull *et al.*, 2009; Pakull *et al.*, 2011; Tuskan *et al.*, 2012; Kersten *et al.*, 2014; Geraldès *et al.*, 2015). Also in *Populus*, the SDR has been mapped to two different positions on chromosome XIX depending on the species (Kersten *et al.*, 2014; Geraldès *et al.*, 2015). In willow species examined thus far, *S. viminalis* and *S. suchowensis*, sex was found to be determined by a ZW system (Semerikov *et al.*, 2003; Hou *et al.*, 2015; Pucholt *et al.*, 2015; Chen *et al.*, 2016) with the SDR mapped to chromosome XV (Hou *et al.*, 2015). Thus, the SDR has translocated during the recent evolution of *Populus* and *Salix*, and the mechanisms of sex determination have also diverged.

Comparative genomic, molecular, and phylogenetic analyses are needed to elucidate the patterns of sex chromosome evolution and how this drives subsequent sexual dimorphism and sex ratio bias. In the strictest sense, primary sexual dimorphism involves distinct morphological features in gamete production. Secondary dimorphism includes all other differences in characteristics between males and females, including morphology, physiology, and phenology (Charlesworth, 1999; Dawson and Geber, 1999; Delph, 1999). Divergent ecological and sexual selection is hypothesized to result from different fitness optima and physiological trade-offs due to the inequality in the energy cost of reproduction (i.e. seed versus pollen production) (Lewis, 1942; Arnold, 1994; Delph, 1999; Obeso, 2002). It has been documented that females of woody dioecious plants typically produce less biomass due to slower vegetative growth as a result of greater allocation to reproduction compared to males (Lewis, 1942; Lloyd and Webb, 1977; Obeso, 2002). Phenotypic traits such as primary growth, production of secondary metabolites, and water use efficiency may be influenced by carbon resource allocation related to sex. Reports

examining sexual dimorphism in *Salix* have shown contrasting results. In a study of *S. sachalinensis* (syn *S. udensis*), no differences were detected in growth or mortality rates between males and females measured in a natural population over a three-year period (Ueno *et al.*, 2007). Conversely, it was reported that drought tolerance and gas exchange rates differed between sexes in *S. glauca*, indicating dimorphism in physiological responses to abiotic stress with lower stomatal conductance (g_s) and transpiration rates in males than in females when exposed to the same drought conditions (Dudley and Galen, 2007). These contrasting studies demonstrates the lack of understanding of the basis for sex dimorphism, particularly considering the limited number of dimorphic characteristics observed.

Another aspect of dimorphism in dioecious plants is sex ratio bias. Classical theories for sex ratios predict a 1:1 ratio of male:female if the expense of resources is the same for producing each sex (Darwin, 1877; Fisher, 1930; Edwards, 2000). New theories suggest that sex ratio bias in natural populations (Delph, 1999; Obeso, 2002) could be due to variation in pollen and seed dispersal (Lloyd, 1982), ecological factors (Barrett *et al.*, 2010), as well as bias in the sex-determination systems (Charlesworth, 2015) and genetic distorters (Taylor, 1999). Since dioecious species are typically perennials (Field *et al.*, 2013a) and can reproduce clonally and sexually, the degree and frequency of flowering and clonal propagation can also influence sex ratios. Several studies examining this in natural populations of *Salix* have shown a female sex ratio bias (Ueno *et al.*, 2007; Che-Castaldo *et al.*, 2015), but this can switch to male bias as demonstrated in a controlled cross experiment with *S. viminalis* (Alström-Rapaport *et al.*, 1997).

How and why sexual dimorphism occurs is not well understood. In general, sexual dimorphism in plants is less developed than in animal systems, and the evidence for dimorphism in secondary characteristics is scarce. The short generation time and greater diversity of *Salix*

relative to *Populus* makes it an interesting system for the study of sexually dimorphic phenotypes and the genetics and evolution of sex chromosomes. This study examined *S. purpurea* L. (purple osier willow), a naturalized species in North America. It is also a model species for willow genomics and an important species in breeding shrub willow bioenergy crops in North America, as it has been used in over 30% of all intra- and interspecific hybrids produced to date (Smart and Cameron, 2008). Critical traits to study for dimorphism include pest and disease resistance, drought tolerance, nitrogen and water use efficiency, and yield. Sexually dimorphic differences in these traits could lead to natural selection and continued evolution of the SDR.

To date, there have been no reports analyzing phenotypic variation or dimorphism in secondary sex characteristics in *S. purpurea*. In this study, I investigated the phenotypic diversity of a natural collection of *S. purpurea* accessions, as well as F₁ and F₂ families produced through controlled cross pollinations. The objectives of this study were to (1) evaluate the phenotypic variation among natural accessions and within breeding pedigrees of *S. purpurea* (2) determine if there is evidence of sexual dimorphism of secondary sex characteristics within natural and bred populations, and (3) test if observed sex ratios fit those expected based on the sex determination system or if there is a sex ratio bias within the species.

MATERIALS AND METHODS

Germplasm and Field Trials

Three populations of *S. purpurea* L. were used in this study: a diverse collection of accessions from the northeast US, an F₁ family produced by crossing two natural accessions, and an F₂ family generated by crossing two F₁ progeny. The diverse collection of accessions included 110 genotypes of *S. purpurea*, where 94 were natural accessions (Lin *et al.*, 2009) and 16 were

horticultural cultivars. For comparison, 18 additional genotypes representing related species and hybrids were also examined (Table A3.1). Three common garden trials were established for the diverse *S. purpurea* collection in July 2012, and all genotypes were hand planted using 20-cm cuttings at three experimental sites (Table A3.2): Cornell University's New York State Agricultural Experiment Station (NYSAES) in Geneva, NY; Cornell University's Lake Erie Research and Extension Lab (CLEREL) in Portland, NY; and the West Virginia University Agronomy Farm in Morgantown, WV. All sites were planted in a randomized complete block design with six replicates of four-plant plots at each location in single-row spacing with 1.82 m between rows and 0.40 m between plants within rows. Border rows containing either genotype 94006 or cultivar 'Fish Creek' were planted on the perimeter to avoid edge effects. At the end of the establishment year, all plants were coppiced and trials were measured in 2013 and 2014 and subsequently harvested and weighed in early 2015. Prior to re-growth of the second rotation in 2015, 112 kg ha⁻¹ N-P-K fertilizer was applied to half of the replicates at each location to test for nitrogen utilization.

The intraspecific F₁ *S. purpurea* family was generated from a cross between the female genotype 94006 and the male genotype 94001, which were accessions collected near Syracuse, NY and were also present in the diverse collection. Two F₁ progeny from this family were selected and crossed ('Wolcott' × 'Fish Creek') to generate the F₂ family. The F₁ and F₂ families and the parents were planted in a single trial. A total of 100 F₁ and 482 F₂ progeny and their parents were hand planted using 20-cm cuttings at Cornell University's New York State Agricultural Experiment Station (Geneva, NY) in June 2014 (Table A3.2) in a randomized complete block design with four replicate blocks of three-plant plots in the same single-row spacing described above. To avoid edge-effects, border rows containing 94006 and 'Fish Creek'

were planted along the perimeter of the trial. At the end of the establishment year, all plants were coppiced, fertilized with 112 kg ha⁻¹ N, 67 kg ha⁻¹ P and K, and measurements were collected in 2015.

Phenotyping

The three trials containing the diverse collection were evaluated for 26 biomass, morphological, phenological, physiological and wood composition traits measured as described below in 2013 and 2014. A subset of traits across three sites as well as rust severity on two sites was measured in 2015. Growth measurements in the diverse collection were conducted on the inner two plants of each four-plant plot. The trial containing the F₁ and F₂ populations was evaluated in 2015 (Table 3.1), where the central plant in each three-plant plot was measured. Rust was surveyed by assessing all the plants in each plot. Sex was scored for clonally propagated plants of each genotype growing in nursery beds and was confirmed in field trial plots.

Table 3.1 Phenotypic traits measured in the *S. purpurea* trials

Trait	Abbreviation	Units
<i>Morphology-Biomass</i>		
Mean stem diameter	SDIA	mm
Total stem area	SA	cm ²
Height	HT	m
Internode length	IL	cm
Stem number	SNo	#
Crown diameter	CDIA	cm
Crown form	CFOR	°
Leaf length	LFL	cm
Leaf width	LFW	cm
Leaf area	LFA	cm ²
Leaf perimeter	LFP	cm
Leaf weight	LFWT	g
Specific leaf area	SLA	cm ² g ⁻¹
Survival	SRV	%
Yield	YLD	dry Mg ha ⁻¹
<i>Phenology</i>		
Vegetative phenology	VPH	day of year
Floral phenology	FPH	day of year
<i>Physiology</i>		
August SPAD	AugSPAD	SPAD units
September SPAD	SeptSPAD	SPAD units
Stomatal conductance	g_s	mmol m ⁻² s ⁻¹
Canopy color	RGB	RGB
<i>Chemical-Physical Wood Properties</i>		
Hemicellulose	HEMI	%
Cellulose	CELL	%
Lignin	LIG	%
Ash	ASH	%
Specific gravity	SPGR	g cm ⁻³
<i>Pathology</i>		
Rust severity	RUST	%

During the dormant period after each growing season, diameters (cm) of stems greater than or equal to 5 mm were measured at 30 cm from the base of the plant using Masser Racal 500 digital calipers and stem number was counted for each plant (Masser, Rovaniemi, Finland). Total stem area (cm²) per plant was also calculated using the stem diameter values. Maximum stem height (m) of every plot was recorded using a measuring rod (Crain Enterprises, Inc.,

Mound City, IL), and the mean height was calculated for each plot. In July of each year, internode length (cm) was measured within the middle third of the tallest stem of each plot and the length of five internodes were recorded. Accounting for different phyllotactic patterns, alternate leaves were counted using five alternate buds or leaves from the first designated bud/leaf, whereas opposite leaves or buds were counted as one node. At the end of the second growing season, crown diameter (cm) was measured using modified Hagl f Mantax blue forestry calipers (Hagl f Sweden AB, L ngsele, Sweden). Stool diameters were measured at 15 cm above the soil, which is the average height of a shrub willow harvester. Crown form (branching angle) was calculated by using one-half of the crown diameter measurement and the height at which it was measured (15 cm) to find the angle of the stem branching relative to the soil. Leaf perimeter (cm), maximum leaf length (cm), leaf width (cm) and leaf area (cm²) were measured on mature leaves at mid-canopy level on the tallest stem of each plant per plot using a laser leaf area meter (CI-203 model, CID Bio-Science, Inc., USA). The same measurement leaves were collected, dried at 65 C, and weighed. The dry weight and measured leaf area were used to calculate specific leaf area (SLA) (cm² g⁻¹ dry wt).

Yield of each plot in the three trials containing the diverse *S. purpurea* collection was measured after the second year of post-coppice by harvesting and weighing all four plants in each plot using the Ny Vraa JF192 harvester (Ny Vraa Bioenergy, Tylstrup, Denmark). Chips were collected in a plastic bin mounted on Avery Weigh-Tronix weigh cells (Fairmont, MN), and the total wet weight of the chip biomass of each plot was recorded. A sub-sample of fresh chip biomass (~1 kg) was collected for each plot, weighed after harvest, oven-dried at 65 C to a constant weight, and dry weight recorded to determine moisture content at harvest. The moisture content was then used to estimate plot dry weights from the measured fresh weights. For all

plots, dry biomass yield was calculated and expressed in dry Mg ha⁻¹ based on plot area.

Phenology

Floral and vegetative bud break were observed and scored using a 0-5 rating scale only in the second year of growth due to the absence of floral buds in the first year. The established scale used for phenology ratings was modified from Saska *et al.* (2010). Both floral and vegetative phenology was surveyed once a week for five weeks and was recorded as the day of the year for a given rating that was observed. All observations occurred until all stage 5 scores were recorded for every genotype. For all trials, the sex of each genotype was visually scored and recorded as either male (M), female (F), or hermaphrodite (H).

Physiology

Stomatal conductance (g_s) (mmol m⁻²s⁻¹) was measured on the abaxial side of the leaf with a leaf porometer (SC-1 Leaf Porometer, Decagon, Pullman, WA) on the uppermost fully expanded leaf of the tallest stem of the plant. A non-destructive proxy for leaf nitrogen status was measured with a portable chlorophyll meter (SPAD-502, Minolta Osaka Co., Ltd., Japan) where readings were collected from three leaves along the length of the tallest stem from the upper, middle, and lower canopy levels and averaged for each plot.

Canopy color (RGB-15) in the trial with the of F₁ and F₂ families was determined by plot using aerial images collected with a gimbal-mounted 14 Megapixel F/2.8 140° FOV camera (w/ lens stabilization) of a Phantom 2 Vision+ (DJI, Nanshan District, Shenzhen, China) quadcopter. To account for any variation, three replicate images were taken for each interval at a fixed altitude (37 m) along the length of the field trial (365 m) in late-July 2015. An overlap of each

interval was required to properly interleave them into a single image. Images were lens corrected using the DJI Vision plugin, ordered, and interleaved using Adobe Photoshop CS6 (Adobe Systems Incorporated, San Jose, CA). The resulting interleaved full-field images were converted into separate RGB channels and analysed by plot using a colorimetric scale based on green pixel density in the open-source program ImageJ v1.47 (Rasband, 1997-2016; Schneider *et al.*, 2012). Excluding aisles and border plants, a coordinate grid of the field was developed in order to obtain average pixel density for each plot.

Wood Properties

Physical and chemical wood properties were measured for four replicates in each of the three trials with the diverse *S. purpurea* collection. Stem segment samples were collected in the dormant period after each growing season using sampling methods previously described (Liu *et al.*, 2015) and were stored frozen at -3°C until they were processed. The specific gravity of each sample was measured by volumetric displacement (TST om-06, 2006). In 2014, a modified method of measuring specific gravity was used where the volume of water displaced was weighed for added precision. Following specific gravity determination, stem segments were oven-dried at 65°C to a constant weight and then rough milled to a 5 mm particle size with a Retch SM300 cutting mill (Retch, Haa, Germany) and were further comminuted to a 0.5 mm particle size by fine milling with the IKA MF 10.1 knife mill (IKA, Wilmington, NC) for compositional analysis. Approximately 20 mg of each milled, unextracted stem sample was analyzed with a Thermogravimetric Analyzer Q500 instrument and Universal Analysis 2000 ver. 4.5A software (TA Instruments, New Castle, DE), as previously described (Serapiglia *et al.*, 2009). Hemicellulose, cellulose, lignin, and ash content were then determined as a percentage of total dry biomass for each sample as previously described (Serapiglia *et al.*, 2014b).

Disease Severity Assessment

In September 2015, leaf rust severity was visually scored in two of the three trials with the diverse collection and in the trial with the F₁ and F₂ families. Percent disease severity was scored (0-100%) for each plot based on leaf area infected and degree of defoliation. Leaf shedding typically occurred when a leaf was 50% infected, thus the highest disease severity recorded was 50%.

Statistical analysis

Statistical analyses and figure generation were conducted with SAS[®] version 9.4 (SAS Institute, Cary, NC) and R version 3.2.3 (R Core Development Team). Mixed linear models were used to analyze phenotypic data implemented in SAS[®] version 9.4 with the PROC MIXED statement and with the *lmer* function within the *lme4* package in R (Bates *et al.*, 2015). Statistical significance for all data analysis was detected at $\alpha \leq 0.05$. All dependent variables were tested for homogeneity of variances and normality using PROC UNIVARIATE using Kolmogorov-Smirnov D and Shapiro-Wilk's K statistics. Non-parametric methods were used when parameters that were not normally distributed could not be transformed to meet the assumptions of parametric analyses using Box-Cox powers or log-transformation. Yield data was square-root transformed to meet assumptions of normality. Pearson's product-moment correlations (*r*) were calculated between all traits. To test for statistical differences between phenotypic traits based on sex, a two-tailed Mann-Whitney *U*-test was conducted with hermaphrodite genotypes excluded. To estimate the predictability and relationship of each trait and biomass yield, multiple linear regression was performed with PROC REG using the stepwise regression method. Model

adequacy was checked with a general linear model to assess the global significance with PROC GLM.

RESULTS

*Phenotypic variation in the diverse *S. purpurea* collection*

For all traits, there were large and significant differences among the genotypes in the diverse *S. purpurea* collection (Table A3.3). Genotype, location and genotype \times location effects for yield were highly significant ($P < 0.05$) (Table 3.2). Mean biomass yield was significantly different among the three sites. Geneva was the most productive location with a site mean of 11.62 Mg ha⁻¹ compared to 6.40 and 9.67 Mg ha⁻¹ for Portland and Morgantown, respectively. The greatest mean yield observed among all locations was produced by commercial cultivars included in the trials with the greatest yield of 57.1 Mg ha⁻¹, which was from the triploid female hybrid *S. viminalis* \times *S. miyabeana* ‘Preble’ at Geneva. When considering only the *S. purpurea* accessions, the greatest yield was 22.5 Mg ha⁻¹ by the female clone 00-01-009 at Geneva, which was 40% of that produced by the best performing commercial hybrid. The greatest *S. purpurea* yields from Portland and Morgantown were 19.9 (05-OSU-063) and 20.6 Mg ha⁻¹ (Pur12), female and male genotypes, respectively. The lowest yielding genotype (03-01-013, female) produced 1.25 Mg ha⁻¹ which was consistent across all locations.

Table 3.2 Mixed model results testing for genotype and locational effects on yield

Source	df	<i>F</i> Ratio	Pr> <i>F</i>
Location	2	213.68	<0.0001*
Genotype	106	10.79	<0.0001*
Genotype x Location	212	1.53	<0.0001*

*Significant differences at $P < 0.05$

Overall, the greatest differences in growth traits between genotypes were for total stem area (SA) and stem height (HT). The range of values across two growing seasons in the three trials were 0.07 to 84.4 cm² for SA and 0.11 to 4.88 m in HT (Table A3.3). There was wide variability in stem number (SNo), which increased on average by 27% from the first to the second year. Crown form (CFOR), calculated from crown diameter (CDIA), ranged from approximately 4 to 88° mean branching angle, but all genotypes had variable CFOR across sites. Of the four metrics obtained from leaf scans, the greatest variation was for leaf perimeter (LFP). The same degree of variability was observed across sites for phenology and physiology traits, where stomatal conductance (g_s) had the greatest variability with the maximum value of 1178.6 mmol m⁻² s⁻¹ and the minimum value of 45.5 mmol m⁻² s⁻¹ in year 1 (Table A3.3, Figure A3.1). The genotypic means for wood composition and specific gravity (SPGR) also had wide variances. The largest variation was observed for cellulose content (CELL) with a range of 20% difference between the highest and lowest value. On average, second year measurements of hemicellulose content (HEMI) and CELL content were greater than first year values by 0.04% and 3.86%, respectively (Table A3.3). Lignin content (LIG) decreased by 1.46% in year 2 compared to year 1 and ash content (ASH) declined by 0.57%, but SPGR only decreased by 0.01 g cm⁻¹ in the second year.

Phenotypic variation in the trial with the F_1 and F_2 families

The growth of the F_1 and F_2 families was on average better than that of the diverse collection in Geneva (Table 3.3, Figure A3.2-A3.3). The means for SA and HT for first-year coppice growth were greater in the F_1 and F_2 families compared to the diverse collection. The first-year mean SA was 16.9 and 12.6 cm² in the F_1 and F_2 families, respectively, while it was 9.32 cm² in the diverse collection in Geneva. Relative differences in first-year post-coppice HT matched those of SA. The mean HT in the F_1 family was 3.26 m and the mean HT of the F_2 family was 3.11 m, while the mean HT in the diverse collection in Geneva was 1.93 m. SPAD measurements and specific leaf area (SLA) showed similar trends with greater means for both traits in the F_1 and F_2 families than in the diverse collection from Geneva.

In general, overall lower trait means were observed for biomass and physiological traits in the F_2 family compared to the F_1 family (Table 3.3). For instance, SDIA, HT, SNo, and SA were all significantly greater ($P<0.01$) in the F_1 family with t -values ranging from 7.3 to 14.7. This was also true for traits related to stem architecture, CDIA ($t=12.6$) and CFOR ($t=12.1$) (Table 3.3). Stem area was ~33% greater in the F_1 family compared to the F_2 family in 2015. Although morphological leaf traits were significantly greater in the F_1 family, SLA and canopy color (RGB) were the only two traits that were not significantly different between F_1 and F_2 families. SPAD was the only trait that was significantly greater in the F_2 family ($t=-3.97$, $P<0.01$).

Table 3.3 Means and standard deviations of phenotypic traits in the *S. purpurea* F₁ family (n=100) and F₂ family (n=482) in Geneva, NY.

	F ₁ <i>S. purpurea</i> family		F ₂ <i>S. purpurea</i> family		
Trait ^a	Mean ± SE	Min – Max	Mean ± SE	Min – Max	<i>t</i> -value ^b
Morphology-Biomass					
SDIA	9.44 ± 0.05	6.36-12.4	8.81 ± 0.02	5.00-12.4	12.9*
SA	16.9 ± 0.28	4.48-33.3	12.6 ± 0.12	2.80-37.7	7.30*
HT	3.26 ± 0.19	2.15-4.78	3.11 ± 0.09	0.98-4.31	14.7*
SNo	21.8 ± 0.32	6.00-41.0	18.7 ± 0.16	1.00-44.0	8.51*
CDIA	36.9 ± 0.43	18.1-84.7	31.3 ± 0.19	3.10-77.0	12.6*
CFOR	40.4 ± 0.32	19.8-59.3	45.2 ± 0.17	21.6-84.2	12.1*
LFL	9.84 ± 0.06	6.52-13.9	9.15 ± 0.03	4.58-18.9	9.97*
LFWD	2.18 ± 0.02	1.40-4.93	2.04 ± 0.01	1.10-12.7	4.90*
LFA	17.2 ± 0.19	8.19-33.8	14.9 ± 0.08	4.08-37.9	11.1*
LFP	22.4 ± 0.28	13.8-57.8	21.0 ± 0.12	9.54-58.2	4.80*
LFWT	0.13 ± 0.002	0.06-0.28	0.11 ± 0.001	0.03-0.41	9.74*
SLA	134 ± 0.89	97.0-233	132 ± 0.42	61.9-361	1.15
SRV	99.7 ± 0.20	33.3-100	99.1 ± 0.15	0.00-100	2.70*
Physiology					
SPAD	55.9 ± 0.29	11.1-76.3	57.2 ± 0.13	26.2-91.9	-3.97*
RGB	112 ± 0.78	55.5-158	111 ± 0.37	48.3-168	0.90
Disease					
RUST	-	-	0.08 ±0.002	0.00-0.86	-

^a Phenotypic traits were measured in 2015. See Materials and Methods for trait definitions and Table 1 for abbreviations and units.

^b Student's *t*-test statistic, where * denotes significant differences among populations at a *P*<0.01 level-of-confidence, with positive values indicating greater means in the F₁ and negative value indicating greater means in the F₂.

Phenotypic analysis of sexual dimorphism

Significant differences by sex were found in *S. purpurea*, with males producing greater growth and significantly greater means for the majority of traits measured (Figure 3.1, Tables 3.4-3.5). In the first year of growth of the diverse *S. purpurea* collection (2013), six traits were significantly dimorphic across three sites ($P < 0.05$, Table 3.4). The primary trait of interest, yield (YLD), was also significantly greater in males. Mean YLD of males ($9.46 \text{ dry Mg ha}^{-1}$) across all three sites was 5.7% greater than females ($8.95 \text{ dry Mg ha}^{-1}$) ($F_{2,1890} = 4.8$, $P = 0.02$). There were significant differences in YLD by site ($F_{2,1890} = 159$, $P < 0.01$), but no significant difference in sex by site interaction ($F_{2,1890} = 0.18$, $P = 0.83$) based on results from the linear mixed model. Yield trends by site were the same as the overall mean comparison, with males producing significantly greater biomass at each location (Figure 3.1). Males had significantly greater means than females for SDIA, internode length (IL), and SPAD measurements from two time points, CELL, and SPGR. Four of the six traits that were significantly dimorphic in year one (2013) were also significantly dimorphic in year two (2014), with the addition of male-biased means for YLD, SA, CDIA, CFOR, leaf weight (LFWT), and LIG. Crown form was calculated from CDIA, and showed significantly lower branching angle in males than females reflecting a wider crown diameter in males. Internode length and CELL did not show dimorphism in 2014. Floral (FPH) and vegetative (VPH) phenology measurements showed significantly lower means for males, indicating earlier bud break for males. In the first year of the second rotation (2015) of the diverse collection trials, male means were significantly greater than females for six of the seven traits measured ($P < 0.05$) (Table 3.4).

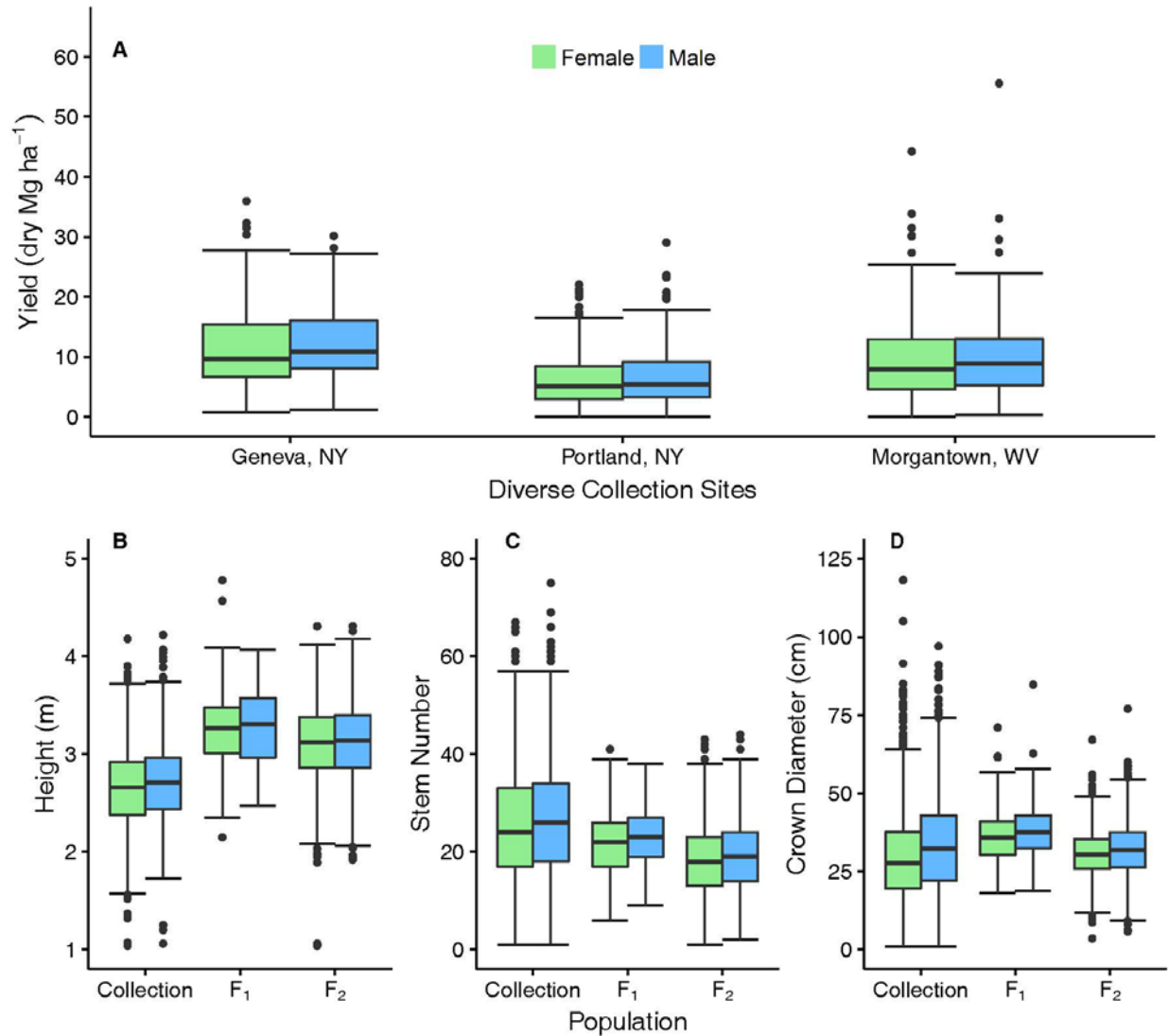


Figure 3.1 Box plots of biomass and morphological traits that were significantly different ($P < 0.05$) between females (green boxes) and males (blue boxes). A) Second year, first rotation biomass yield from the diverse collection trials measured across three field sites. B) Plant stem height measured in 2015 for each population. C) Plant stem number measured in 2015 for each population. D) Plant crown diameter measured for each population. Data from the diverse collection was averaged across three sites.

Table 3.4 Comparison of phenotypic traits for female and male individuals in the diverse collection of *S. purpurea* across three growing seasons.

Trait	Female (n=52)		Male (n=55)		P-value	Dimorphism (%)
	Mean ± SE	CV (%)	Mean ± SE	CV (%)		
2013						
SDIA	7.21 ± 0.05	23.02	7.4 ± 0.05	20.54	0.01*	2.64
SA	9.26 ± 0.23	76.24	9.34 ± 0.22	72.06	0.44	0.86
HT	1.93 ± 0.02	27.98	1.91 ± 0.02	27.75	0.35	-1.04
IL	13.5 ± 0.17	37.63	13.8 ± 0.16	34.93	0.01*	2.22
SNo	18.2 ± 0.33	55.56	18.3 ± 0.33	55.56	0.92	0.00
LFL	6.81 ± 0.07	24.38	6.93 ± 0.06	23.38	0.16	1.76
LFW	1.68 ± 0.05	70.83	1.8 ± 0.06	78.33	0.88	7.14
LFA	8.69 ± 0.17	50.06	8.92 ± 0.18	50.00	0.32	2.65
LFP	23.4 ± 0.95	15.81	24.7 ± 1.11	33.20	0.15	5.56
LFWT	0.075 ± 0.001	42.67	0.077 ± 0.001	54.55	0.17	2.67
SLA	128 ± 2.44	71.02	132 ± 3.2	70.83	0.34	3.13
AugSPAD	45.9 ± 0.3	20.46	47.1 ± 0.3	20.02	0.01*	2.61
SeptSPAD	41.3 ± 0.33	20.15	42.9 ± 0.37	21.86	<0.01*	3.87
g _s	601 ± 6.91	33.11	595 ± 6.53	32.27	0.80	-1.00
HEMI	17.5 ± 0.04	5.26	17.6 ± 0.03	4.94	0.17	0.57
CELL	37.5 ± 0.13	8.53	37.9 ± 0.13	8.71	0.02*	1.07
LIG	28.9 ± 0.09	7.40	28.8 ± 0.08	7.01	0.26	-0.35
ASH	2.15 ± 0.02	26.98	2.13 ± 0.02	28.64	0.63	-0.93
SPGR	0.45 ± 0.003	15.56	0.46 ± 0.003	13.04	0.01*	2.22
2014						
YLD	4.48 ± 0.06	69.87	4.75 ± 0.15	62.95	<0.01*	6.03
SDIA	9.85 ± 0.07	20.41	10.1 ± 0.06	19.31	<0.01*	2.54
SA	20.7 ± 0.45	66.67	21.8 ± 0.43	61.01	0.02*	5.31
HT	3.13 ± 0.02	21.73	3.07 ± 0.02	21.82	0.06	-1.92
IL	13.6 ± 0.23	42.06	13.4 ± 0.22	41.27	0.86	-1.47
SNo	22.4 ± 0.38	54.55	23.0 ± 0.37	47.83	0.14	4.55
CDIA	29.6 ± 0.5	51.35	34.2 ± 0.75	68.42	<0.01*	15.54
CFOR	49 ± 0.45	28.16	45.8 ± 0.47	32.10	<0.01*	-6.53
LFL	6.53 ± 0.06	29.25	6.66 ± 0.06	26.88	0.23	1.99
LFW	1.8 ± 0.06	95.56	1.79 ± 0.06	96.65	0.98	-0.56
LFA	8.61 ± 0.13	47.74	8.7 ± 0.12	43.56	0.42	1.05
LFP	17.4 ± 0.29	51.38	17.6 ± 0.27	47.33	0.28	1.15
LFWT	0.09 ± 0.001	44.44	0.1 ± 0.001	40.00	<0.01*	11.11
SLA	92.3 ± 0.8	26.54	89.5 ± 0.66	23.13	0.01*	-3.03
VPH	111 ± 0.14	3.08	108 ± 0.15	3.57	<0.01*	-2.70
FPH	95.9 ± 0.53	13.87	87.1 ± 0.59	17.22	<0.01*	-9.18
AugSPAD	44.9 ± 0.23	12.74	47.3 ± 0.22	11.80	<0.01*	5.35

Table 3.4 (Continued)

SeptSPAD	42.5 ± 0.27	18.14	44.8 ± 0.26	16.83	<0.01*	5.41
<i>g_s</i>	480 ± 5.58	35.63	492 ± 5.64	35.77	0.10	2.50
HEMI	17.7 ± 0.03	4.29	17.6 ± 0.03	4.20	0.06	-0.56
CELL	41.6 ± 0.07	4.25	41.7 ± 0.07	4.03	0.67	0.24
LIG	27.3 ± 0.05	4.40	27.5 ± 0.04	4.15	0.01*	0.73
ASH	1.55 ± 0.02	24.52	1.58 ± 0.02	24.05	0.18	1.94
SPGR	0.49 ± 0.001	10.20	0.5 ± 0.001	8.00	<0.01*	2.04
2015						
SDIA	7.46 ± 0.04	16.22	7.63 ± 0.03	14.15	<0.01*	2.28
Table 3.4 (Continued)						
SA	12.3 ± 0.23	57.32	13.4 ± 0.24	56.42	<0.01*	8.94
HT	2.64 ± 0.01	16.29	2.7 ± 0.01	14.81	<0.01*	2.27
SNo	25.1 ± 0.38	44.00	27.5 ± 0.39	44.44	<0.01*	8.00
LFA	13.7 ± 0.2	36.28	14.1 ± 0.18	33.33	0.15	2.92
AugSPAD	48 ± 0.27	17.10	50 ± 0.26	16.28	<0.01*	4.17
RUST	28.1 ± 0.49	44.13	29.5 ± 0.54	46.78	0.03*	4.98

Values are mean ± SE across three locations. Two-tailed Mann-Whitney *U*-test (df=1) results, where significant values ($P<0.05$) are denoted by bold font and *. Positive values for dimorphism denote male-biased difference and negative values denote female-biased difference.

A specific aim in the second rotation was to measure differences in nitrogen utilization after fertilizing half of the replicate blocks at each site. Significant differences ($P<0.01$) (Table 3.6) were found by site, treatment, sex and site by treatment interaction, but no significant sex by treatment interaction ($P=0.58$) (Table 3.6). At each location the fertilized plots had greater SPAD values than the controls regardless of sex. Additionally, males had higher SPAD values than females in both treated and control plots at each location (Figure 3.2). Leaf area was not significantly different between sexes ($P=0.15$), but still exhibited ~3% greater trend for male genotypes.

Four traits were sexually dimorphic in the trial with the F_1 and F_2 families (Table 3.5). Male means for HT, CDIA, and SLA were greater than females, while CFOR was greater in females. In the F_1 family, SDIA of female progeny was greater than that of males, whereas SNo

was greater in males compared to females. In the F₁ family, leaf weight (LFWT), LFP, and leaf length (LFL) means were greater in females than males.

Table 3.5 Comparison of phenotypic traits for male and female individuals in a F₁ *S. purpurea* family (n=100) and a F₂ *S. purpurea* family (n=482) measured in 2015 in Geneva, NY.

Trait	F ₁ <i>S. purpurea</i> family						F ₂ <i>S. purpurea</i> family					
	Female (n=70)		Male (n=30)		P-value	Dimorphism (%)	Female (n=266)		Male (n=216)		P-value	Dimorphism (%)
	Mean ± SE	CV (%)	Mean ± SE	CV (%)			Mean ± SE	CV (%)	Mean ± SE	CV (%)		
SDIA	9.52 ± 0.06	9.87	9.26 ± 0.09	10.91	<0.01*	-2.73	8.82 ± 0.03	9.75	8.81 ± 0.03	10.67	0.43	-0.11
SA	16.8 ± 0.34	33.81	17.1 ± 0.49	31.05	0.22	1.79	12.4 ± 0.16	41.69	12.9 ± 0.19	43.64	0.12	4.03
HT	3.24 ± 0.22	11.42	3.29 ± 0.34	10.94	0.02*	1.54	3.11 ± 1.15	12.22	3.13 ± 1.35	12.78	0.02*	0.64
SNo	21.2 ± 0.4	33.33	23.3 ± 0.53	26.09	0.03*	9.52	18.4 ± 0.21	38.89	19.0 ± 0.25	36.84	0.11	5.56
CDIA	36.3 ± 0.49	22.59	38.7 ± 0.88	24.44	0.02*	6.61	30.6 ± 0.24	25.23	32.1 ± 0.3	27.32	<0.01*	4.90
CFOR	40.9 ± 0.38	15.45	39.1 ± 0.6	16.47	0.02*	-4.40	45.8 ± 0.22	16.05	44.6 ± 0.27	17.96	<0.01*	-2.62
LFL	9.84 ± 0.07	11.99	9.83 ± 0.12	12.82	0.85	-0.10	9.19 ± 0.04	13.38	9.09 ± 0.04	14.52	0.03*	-1.09
LFWD	2.19 ± 0.02	15.98	2.16 ± 0.04	18.98	0.32	-1.37	2.05 ± 0.01	21.95	2.04 ± 0.02	27.94	0.13	-0.49
LFA	17.2 ± 0.23	22.33	17.1 ± 0.4	25.26	0.56	-0.58	15.1 ± 0.11	23.71	14.8 ± 0.12	24.73	0.07	-1.99
LFP	22.2 ± 0.33	24.86	22.7 ± 0.53	25.07	0.51	2.25	21.1 ± 0.16	25.21	20.8 ± 0.18	26.11	0.02*	-1.42
LFWT	0.13 ± 0.002	23.08	0.13 ± 0.003	30.77	0.82	0.00	0.12 ± 0.0001	25.00	0.11 ± 0.001	27.27	<0.01*	-8.33
SLA	131 ± 1.07	13.82	134 ± 1.58	14.33	<0.01*	2.29	131 ± 0.55	13.82	134 ± 0.65	14.33	<0.01*	2.29
SRV	99.6 ± 0.26	4.44	99.7 ± 0.29	3.10	0.87	0.10	99.34 ± 0.17	5.51	98.8 ± 0.25	7.43	0.05	-0.54
SPAD	56.2 ± 0.3	8.52	54.9 ± 0.68	12.28	0.11	-2.31	57.2 ± 0.19	10.05	57.1 ± 0.19	8.88	0.88	-0.17
RGB	112 ± 0.94	14.11	111 ± 1.42	13.78	0.38	-0.89	111 ± 0.49	14.50	112 ± 0.55	14.46	0.08	0.90
RUST	-	-	-	-	-	-	7.9 ± 0.002	88.61	8.8 ± 0.003	90.91	0.03*	11.39

Values are mean ± SE. Two-tailed Mann-Whitney *U*-test (df=1) results, where significant values ($P<0.05$) are denoted by bold font and *. Positive values for dimorphism denote male-biased difference and negative values denote female-biased difference.

Table 3.6 Mixed model test for nitrogen utilization

Source	df	<i>F</i> Ratio	Pr> <i>F</i>
Location	2	65.26	< 0.0001 *
Treatment	1	170.00	< 0.0001 *
Sex	1	15.75	0.0001 *
Sex x Treatment	1	0.31	0.58
Location x Treatment	2	6.51	< 0.01 *

*Significant differences at $P<0.05$

Leaf rust severity (RUST) was measured during the 2015 growing season in two of the trials with the diverse collection and in the F₂ family. RUST severity scores were significantly higher ($P<0.05$) for males than females (Figure 3.3). There were significant differences between the two sites surveyed ($P<0.05$) for the diverse collection. Based on disease severity using least square means, males had a higher mean score (29%) for RUST severity than females (26%) in Geneva, NY. However, there was no significant difference in severity by sex at the Portland, NY site (Figure 3.3A).

The male parents of the F₁ and F₂ families had significantly greater mean RUST scores than the female parents (Figure 3.3B) of the families in the trial in Geneva, NY. Similarly, the male F₂ progeny had significantly greater mean RUST severity than the F₂ female progeny ($P=0.02$). The overall F₂ progeny means for RUST severity were greater than that of the female parent, ‘Wolcott’, but less than that of the male parent ‘Fish Creek’. Overall, there was a significant negative correlation between RUST severity and both SA and HT ($P<0.05$), with a significant positive correlation between RUST severity and SPAD measurements ($P<0.01$) (Figure A3.4).

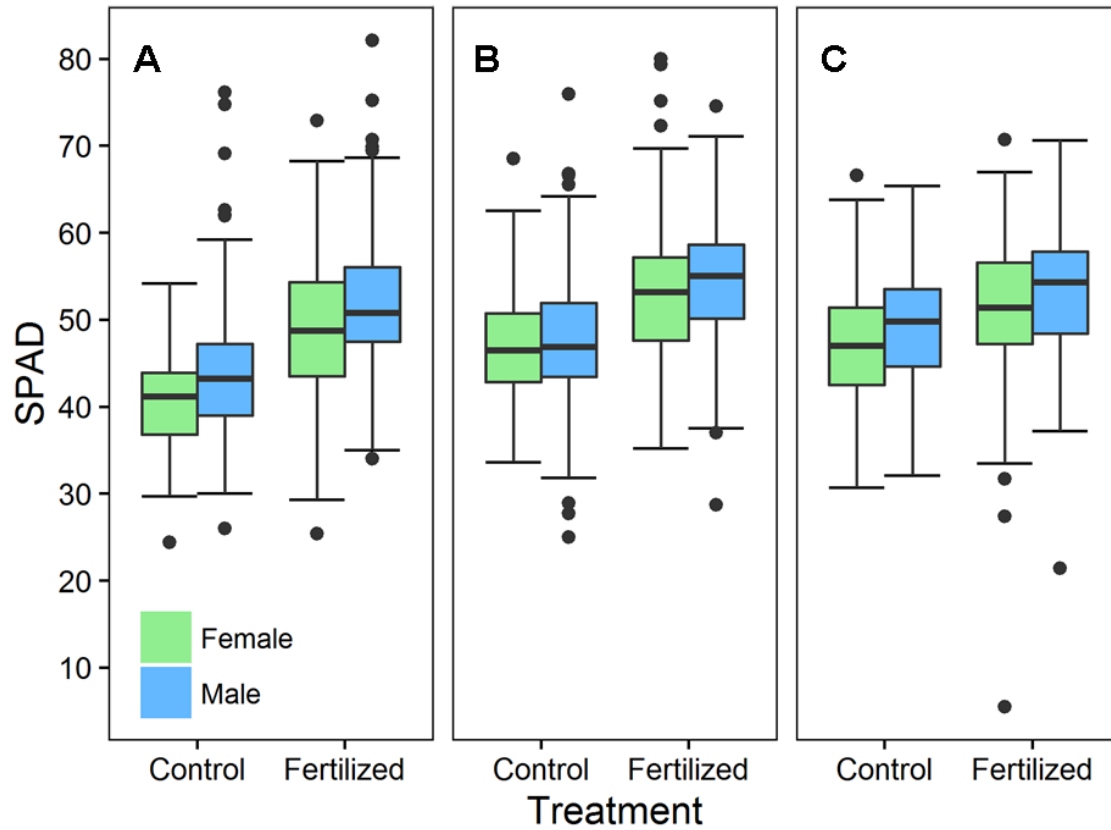


Figure 3.2 SPAD values for monitoring nitrogen utilization in diverse *Salix purpurea* collections. Box plots representing females are colored in green and males colored in blue. SPAD values for control and fertilized plots for A) Geneva, NY ($F_{1,105}=15.73$, $P<0.01$), B) Portland, NY ($F_{1,105}=3.44$, $P=0.06$), C) Morgantown, WV ($F_{1,105}=5.96$, $P=0.02$).

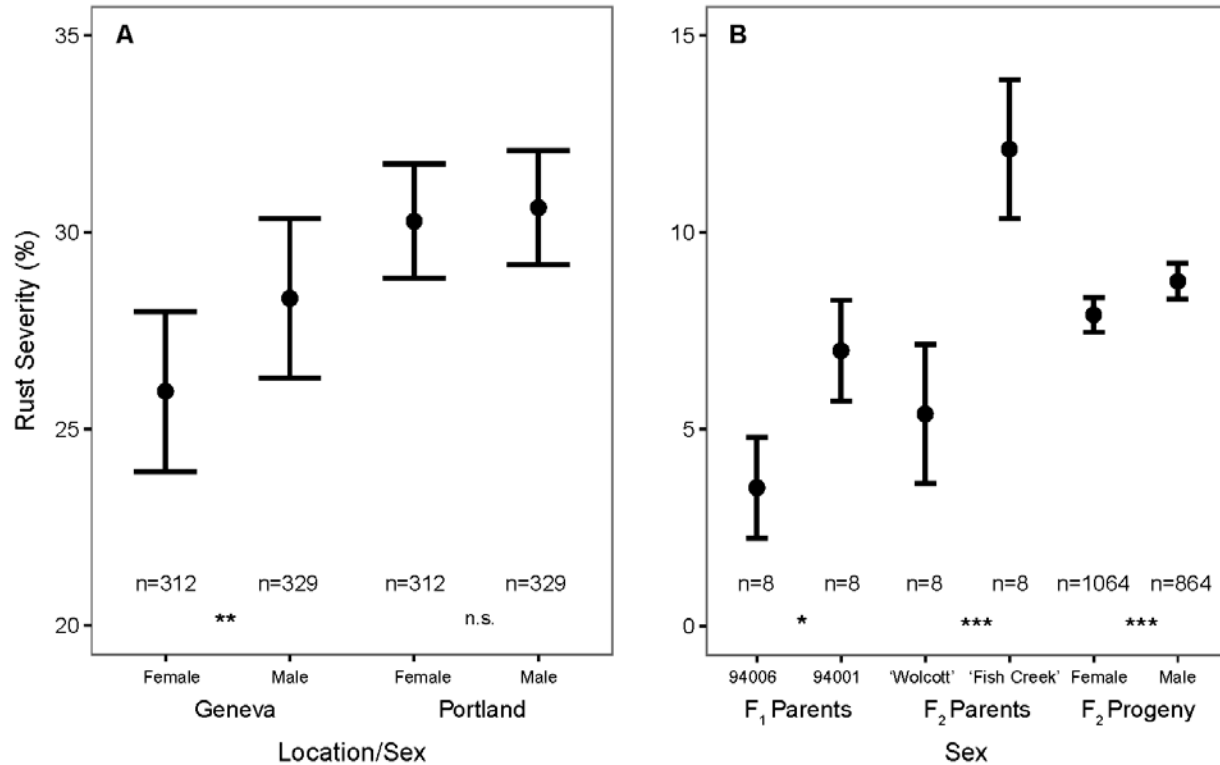


Figure 3.3 Least square means for leaf rust severity scores of female and male *S. purpurea*. A) Rust severity scores and standard errors on females and males of the diverse collection at the Geneva, NY and Portland, NY field sites. B) Rust severity scores and standard error on F₁, F₂ parents and female and male F₂ progeny in Geneva, NY. Significant difference between females and males within each site and population are denoted by *significant at 0.10, **significant at 0.05, ***significant at 0.01, and n.s. for no significant difference.

Sex ratios

The sexes of the genotypes in the diverse collection were 55 male, 52 female, and three hermaphrodite, which were confirmed across years and experimental locations based on documented sex phenotypes in nursey beds. The F₁ family consisted of 70 females and 30 males and the F₂ family contained 266 females and 216 male genotypes. There was significant departure from the expected 1:1 segregation ratio of males and females in both the F₁ and F₂ families. The F₁ family had a female to male ratio of 2.33:1 ($P < 0.01$), and the F₂ family had a significantly female-biased sex ratio of 1.23:1 ($P = 0.02$), but there was no significant departure

from a 1:1 ratio among the genotypes in the diverse collection.

Allometric model for yield

All measured traits from the trials containing the diverse *S. purpurea* collection were used as parameters in allometric models to identify relationships between YLD and the yearly growth measurements using multiple linear regression to predict second year biomass. Separating genotypes by sex and examining allometric relationships with YLD revealed no significant difference ($P=0.15$) or advantage of predicting YLD and therefore the data were not distinguished by sex in the regression model. Variable inflation factors greater than 10 were observed between total SDIA and total SA and indicated multicollinearity as indicated with the high correlation coefficient of $r=0.95$ (Figure A3.5). Since SA had a greater correlation with YLD and explained a greater percentage of the variance in the model, it was kept and SDIA was removed. All other variables that did not meet the $P<0.05$ significance level were also removed. To test for global significance of variables, a general linear model was fitted and revealed SLA in 2014 ($P=0.88$) and LFP ($P=0.84$) were insignificant and were also removed as predictor variables. The best predictors for the final multiple linear regression model were SA in 2013 and 2014, HT in 2014, and AugSPAD in 2014 (Table A3.4), where yearly SA measurements gave the most accurate estimates of YLD. A strong positive fit of predicted and observed biomass YLD resulted in an overall $R^2 = 0.79$ (Figure A3.6).

DISCUSSION

For over a century, botanists, including Darwin, have presented theories to explain the conditions and factors that favor or contribute to the evolution of dioecy (Darwin, 1877; Lewis,

1942; Westergaard, 1958; Lloyd and Webb, 1977; Bawa, 1980; Charlesworth and Guttman, 1999; Delph, 2009). Sexual dimorphism for secondary traits that provides differential expression in one sex over the other could contribute to the maintenance of dioecy over time. For dioecious species, genetic factors, selection pressures over time, and ecological adaptation can cause differential expression of certain phenotypic traits to support dissimilarities in reproductive fitness (Sakai and Weller, 1999). Of the publications that report on dioecious species, very few provide evidence of sex ratio bias, but for those that do, there tends to be a greater frequency of male-biased sex ratios (Field *et al.*, 2013a). The primary interests of the current study was to test whether sexual dimorphism and sex ratio bias exists in the bioenergy crop, *S. purpurea*, across unrelated and full-sib populations in experimental trials.

Many dioecious plants exhibit biased population sex ratios. The most well documented cases of female-biased sex ratios are within *Silene*, *Rumex*, *Cannabis*, and *Humulus* (Lloyd, 1974; Taylor, 1999; Stehlik and Barrett, 2006). For *Salix*, the closest taxonomically related genus, *Populus*, has been reported to show male-biased ratios (Tuskan *et al.* 2012). Sex ratio bias has been reported in *Salix* spp., and is most often biased towards females in an approximate 2:1 (f:m) ratio (Alström-Rapaport *et al.*, 1997; Rottenberg, 1998; Dudley, 2006; Ueno *et al.*, 2007; Hughes *et al.*, 2009; Myers-Smith and Hik, 2012). Even though some studies of experimental populations (Mosseler and Zsuffa, 1989) have revealed greater variability of sex ratio bias than in natural populations (Myers-Smith and Hik, 2012), sex ratio bias in progeny of controlled crosses is dependent on the nature of the cross (inter- or intraspecific) as well as the ploidy levels of the parents. Male ratio bias is commonly seen in trees and often associated with biotic pollen dispersal in contrast to female-biased ratios observed in shrubs and herbs that tend to be clonal perennial species (Field *et al.*, 2013b, 2013a). The female bias observed in the F₁ and F₂

families of this study may be explained by the occurrence of pollen certation or a breakdown in the mechanisms controlling sex determination (Charlesworth, 2002). Especially when pollen load is high, as in the case with controlled crosses, the female determining pollen may be inherently more successful at fertilization. In natural stands of dioecious species, this may have evolved as a useful mechanism to adjust population sex ratios when there was an overabundance of pollen coming from a predominate stand of plants with staminate flowers (Lewis, 1942). This may also explain the expected 1:1 f:m ratio in the diverse *S. purpurea* collection, which originated from natural stands of open-pollinated plants under what was likely a low density of male plants. I can speculate that reduction of the f:m ratio from the F₁ to F₂ generation is a result of slight inbreeding depression, differential mortality, or environmental factors. It should be noted that the hermaphrodites discovered in the diverse collection were not observed to be diphasic during the years under study; they were continually hermaphroditic under all observations. This evidence substantiates predominately genetic and non-environmental control of sex determination by a single locus, which has been proposed in both *S. suchowensis* and *S. viminalis* (Hou *et al.*, 2015; Pucholt *et al.*, 2015).

Despite the female-biased sex ratios in the F₁ and F₂ families, these results clearly demonstrated males were superior to females for most traits across four field trials and three populations of *S. purpurea*. Comparison of coefficients of variation (CV) between sexes in each population showed consistently greater variation in males, which may indicate greater plasticity in response to environmental conditions. The lower CV of female plants may be restricted to the fact that more resources are spent on seed production and variable responses to growth cannot be afforded. Among the natural accessions, every significant difference in trait means for both growing seasons, except for SLA in 2014 was greater in males than females, where AugSPAD,

and SeptSPAD measurements showed consistently higher values in males across years. Traits that had a positive correlation with YLD, also showed greater trait means in males. Direct measurement of YLD was significantly greater in males as well. Within the F₁ family, all traits that were dimorphic were male-biased, except for SDIA. Vegetative and floral bud break are also important traits that can determine the effective length of the growing season. It was observed that males have earlier vegetative bud break, which would extend the growing season for males and may partly explain differences in yield. It has been shown that light use efficiency is an important factor contributing to aboveground biomass and that early bud break may contribute to this production (Tharakan *et al.*, 2008). However, it has been suggested that other phenological events (i.e. leaf unfolding and duration, growth cessation and leaf abscission) affect annual biomass production, where growth cessation and late season leaf retention may also impact aboveground growth as well as nutrient recycling and storage *in planta* (Weih, 2009). An intensive study monitoring these additional traits could provide clues as to whether sex-specific physiological patterns are seen across other dioecious species as well.

Sex dimorphism has also been studied extensively in the closely-related genus *Populus*. Examination of phenotypic and gene expression data in *P. tremula* has shown no evidence of sexual dimorphism (Robinson *et al.*, 2014) for morphological or biochemical traits. Minor differences across a set of growth traits observed in *P. euphratica* suggested that male trees are more vigorous than females (Petzold *et al.*, 2012). Studies of *P. deltoides* and *P. tremuloides* hybrids revealed significantly greater biomass production in males (Pauley, 1948; Farmer, 1964). When examining the evidence for sexual dimorphism in *Salix*, there are just as many opposing observations. In *S. planifolia*, it was reported that a larger allocation of resources are needed for reproduction in females than in males (Turcotte and Houle, 2001), which may lead to the

assumption that females exhibit less biomass growth compared to males. Other studies have shown that females have growth rates similar and sometimes greater but not significantly different than males (Åhman, 1997; Sakai *et al.*, 2006).

These data revealed consistent trends of dimorphism for CDIA and the associated CFOR showed a significant male bias for greater CDIA and subsequent shallower branching angle for all years and populations. Specific gravity contributes to the mechanical properties of wood and is known to scale positively with biomechanical strength and therefore directly influences plant architecture (Chave *et al.*, 2009). This has practical importance for shrub willow, because plants with a wide crown diameter that expands into the alleys of a field will result in biomass that may not be collected by a harvester. If specific cultivars have wide branching angles, this will result in a loss of harvestable biomass and reduction in yield. It has been observed that tree species with greater wood density also have greater horizontal branch expansion and wider crowns (Iida *et al.*, 2012). Male plants in this study had significantly greater SPGR and CDIA, where the opposite trend was seen in females. This suggests that the biomechanical differences imposed by wood density may contribute to the differences in plant architecture and may be dependent on plant sex especially in woody dioecious species.

Another interesting similarity of sexual dimorphism observed throughout this study was the significantly higher rust severity on male plants. *Melampsora* leaf rust is the most severe plant disease affecting short-rotation willow plantations where long-term stability of yield will depend on host resistance. Resistance to *Melampsora* spp. has been mapped in *Salix* and is currently an ongoing effort (Rönnerberg-Wästljung *et al.*, 2008; Hanley *et al.*, 2011; Samils *et al.*, 2011). Among the studies of rust severity in willow, there are few that have provided information on sex dimorphism. Literature surveys conducted to investigate this topic were

mostly surveys among largely unrelated commercial cultivars (Moritz *et al.*, 2016), but these showed greater rust severity on female plants. Only two studies showed male-biased infection (McCracken and Dawson, 2003; Pei *et al.*, 2008). This study found significantly greater rust severity on males consistent in the diverse collection and in the F₂ family for which there should be sufficient genotypic diversity. Previous studies using a polyculture approach reduced overall infection severity due to greater clonal diversity (Åhman, 1997; Begley *et al.*, 2009) suggesting that large interclonal variation might outweigh any sex-specific effects in a field planting. Although, no significant differences were observed in rust severity between males and females at the Portland, NY site, this could be due to timing of RUST assessment. This site was surveyed late in the season and disease was already advanced causing extensive pre-mature defoliation and rendering phenotypic differences that may exist between sexes indistinguishable. This suggests that differential rust susceptibility could also be driven by phenological differences between male and female plants. It is hypothesized that through life-history trade-offs of sexual morphs in dioecious species, females typically allocate greater resources towards reproduction and defense against pests and diseases (Seger and Eckhart, 1996; Vega-Frutis *et al.*, 2013), and males invest more resources into primary growth (Delph, 1999; Obeso, 2002). Additionally, there may be differences in mechanical or biochemical defense mechanisms that were not measured and for which there are limited studies examining this topic (Bañuelos *et al.*, 2004). A possible explanation of why greater rust severity was observed in males may be related to the nitrogen amendments applied to the association trials prior to the beginning of the second rotation. The SPAD values observed in this study, used as a non-destructive method to quantify nitrogen status in the plant, showed significantly greater values in treated versus control plots, but also significantly greater values in males than females in the diverse collection. This suggests that

males may have a greater capacity for nitrogen utilization possibly due to the nitrogen requirement of pollen production (Carolyn and Rundel, 1979) and the greater resource allocation towards primary growth. A review conducted by Hultine (2016) examining differential resource acquisition between sexes of 22 species across multiple environments, concluded that females generally do not have greater nutrient uptake or efficiency over males under optimal growing conditions. Based on these reasons, the results make biological sense for males to exhibit greater rust severity because they would serve as a better host for the obligate biotroph *Melampsora* spp. (Kenaley *et al.*, 2014).

Although a strong male dominant bias was observed, there were also female-biased results as well as traits with no dimorphic differences observed, especially within the F₁ and F₂ families. The selection of parents used to generate these families may have influenced some of these observations. Additional comparisons of the F₁ and F₂ families beyond sexual dimorphism, reveals some evidence of inbreeding depression. For the majority of the traits measured, especially biomass related characters, there was an overall decrease in trait means in the F₂ family compared to the F₁ family. These reductions in biomass-related traits in subsequent generations is likely due to greater homozygosity in this obligate outcrossing species (Charlesworth and Willis, 2009). Similar patterns have been observed in *S. viminalis* (Rönnberg-Wästljung, 2001), while some full-sib F₂ families of *S. eriocephala* have shown inbreeding depression (Aravanopoulos and Zsuffa, 1998) and others have not (Phillips, 2002). There is strong evidence of heterosis in F₁ species hybrids of *Salix* (Serapiglia *et al.*, 2013; Serapiglia *et al.*, 2014a; Serapiglia *et al.*, 2014b), but these pedigrees have not been advanced to the F₂ generation.

CONCLUSION

Significant evidence was found for sexual dimorphism for a majority of traits with a male bias in growth performance, but a female-biased sex ratio in *S. purpurea*. This provides a testable hypothesis that the SDR is linked to loci responsible for growth and fitness traits in *S. purpurea*, which can be pursued in future studies. It is still unclear if sexual dimorphism exists in other *Salix* spp. while studies of *P. trichocarpa* and *P. tremula* have returned evidence of no sex dimorphism. These results should also shed light on the evolution of sex dimorphism in other dioecious plants and the implications of dimorphism on sex ratio bias. Broad comparative analysis across many plant taxa using genomic and phenomic approaches will be necessary to acquire a better understanding of the ecological, evolutionary, and molecular mechanisms controlling sex determination and sexual dimorphism in dioecious plants.

CHAPTER 4 - Genome-Wide-Association Study for a Suite of Bioenergy Traits in Shrub

Willow (*Salix purpurea*)¹

ABSTRACT

Sustainable sources of renewable bioenergy are in demand which requires fast and efficient development of feedstock crops through plant breeding. The aim of this study was to conduct a quantitative and association genetics study in *Salix purpurea* to dissect the genetic regulation of complex traits related to biomass production. A population of 110 individuals was genotyped using GBS and 25,566 SNPs were used to test for associations with 23 phenotypes measured in three replicated field trials across three years while 251 genotypes were used to map the sex determination locus. Marker-based estimations of narrow sense heritability were calculated and were low to moderately high across all traits ($h^2=0.01$ to 0.63). By using three methods of mixed liner models (MLM), 95 significant associations were found for nine phenotypic traits.

Associations reaching genome-wide significance at $p<0.05$ included phenological, physiological traits, leaf rust severity, strong associations for a sex determination locus, and five biomass related traits which included SNPs associated with yield. These results show the potential of GWAS in *Salix* and provide an important foundation for the development of shrub willow bioenergy crops through the advancement of molecular breeding.

¹Chapter 4 is currently being prepared for publication. This work was in collaboration with Luke Evans, Steve DiFazio, Ben Bubner, Matthias Zander, and Larry Smart. My major contribution was designing the study, collecting phenotypic data, carrying out the data analysis, interpretation of the results, and drafting the manuscript.

INTRODUCTION

Long-lived woody perennials such as trees and shrubs have proven to be reliable lignocellulosic feedstocks for second generation biofuel production (Smart and Cameron, 2008; Sannigrahi *et al.*, 2010; Hanley and Karp, 2013). The use of biomass crops under short rotation coppicing or short rotation forestry systems provides fast growth and high yields with relatively low agricultural inputs, which are characteristics that will help mitigate problems with increasing demand for food and energy, with decreasing availability of land and resources (Valentine *et al.*, 2012). It is apparent that significant efforts are needed to develop accelerated crop breeding strategies, and with the advent of high-throughput, next-generation sequencing technologies (Goodwin *et al.*, 2016), and low-cost genotyping protocols, the ability to speed up breeding and selection, especially with non-model species, will continue to improve (Kim *et al.*, 2016) .

Future breeding efforts will most likely follow methods that are robust and fast at dissecting complex traits, such as genome-wide association studies (GWAS) (Soto-Cerda *et al.*, 2013) and genomic prediction methods (genomic selection, GS) (Meuwissen *et al.*, 2001). These molecular breeding approaches have already revolutionized animal and plant breeding and can greatly aid in answering basic biological questions and contribute to applied breeding objectives by narrowing the gap between true biological regulation of phenotypes and the theoretical work of statistical genomics. Marker-assisted selection (MAS) (Collard and Mackill, 2008) for major gene traits would be especially beneficial in perennials with long generation times and breeding cycles (Crossa and Federer 2012), if genotypes with desired characteristics can be selected at the seedling stage, thereby reducing the time and cost required of extensive field trials (Allwright and Taylor, 2016).

Linkage-disequilibrium (LD) based, association mapping (AM) is an alternative approach

to quantitative trait loci (QTL) mapping which uses a set of unrelated genotypes that is designed to capture most of the natural genetic variability for the trait of interest and represents all historic recombination cycles, providing a theoretically higher resolution than QTL mapping (Pritchard *et al.*, 2000b; Mackay *et al.*, 2009). In plant populations of unrelated individuals, the LD is expected to be low due to many historical recombination events and also if the species under study is an obligate outcrosser as is the case for all dioecious species (Khan and Korban, 2012). Generally, AM can be divided into naïve GWAS and candidate gene AM which are influenced by sample size, *a priori* information (gene function), and objectives of the studies (Zhu *et al.*, 2008). The candidate gene approach has been the most ubiquitous in previous studies because of the availability of existing genomic resources for a number of model species, such as *Arabidopsis thaliana*, rice, sorghum, grape, and *Populus trichocarpa* (Meinke *et al.*, 1998; Tuskan *et al.*, 2004) (International Rice Genome Sequencing Project, 2005; Jaillon *et al.*, 2007; Paterson *et al.*, 2009). However, whole-genome association studies have the advantage of assessing the entire genome for trait-associated variants, rather than being limited by the number and specific choice of candidate genes (Gaut and Long, 2003). Additionally, the ability to generate thousands of SNPs genome-wide and generate a reference genome for non-model crops at increasingly lower costs provides greater opportunity for discovery.

Association studies examining various growth, physiological, phenological, and wood composition traits have been conducted in a number of woody perennial species, such as *Eucalyptus urophylla* (Denis *et al.*, 2013), *P. balsamifera* (L.) (Olson *et al.*, 2013), *P. tremula* (Ingvarsson *et al.*, 2008), *P. trichocarpa* (Torr. & Gray) (Evans *et al.*, 2014; McKown *et al.*, 2014), and *P. deltoides* (Fahrenkrog *et al.*, 2016), but so far only one association study has been conducted on willow, which examined *S. viminalis* (Hallingbäck *et al.*, 2015). This is a shrub

species that is bred primarily for bioenergy and is the reference species for several European breeding programs (Karp *et al.*, 2011). The trait associations found in *S. viminalis* were related to various phenology and growth phenotypes, but the study only used 1,536 SNPs and used *Populus* as a reference genome for part of the candidate gene selection process since a reference genome for *S. viminalis* is not yet publically available.

Target phenotypes for mapping studies must also take into consideration the genetic architecture of the trait and how that might influence future breeding efforts for MAS. Early selection of plants with targeted traits would greatly increase the efficiency of breeding, especially if selection accuracy is high. Heritability is an important parameter in quantitative genetics which influences the response to selection and provides an understanding of the proportion of phenotypic variance. For clonally propagated perennial crops like shrub willow, broad-sense heritability (H^2) can be calculated using the ratio of total genetic variance to phenotypic variance. Heritability in the narrow-sense (h^2) only captures the additive genetic variance, but is a useful concept for breeding and selection as it explains the maximum variance by all allelic combinations for a specific trait and permits breeders to maximize genetic improvement.

Previous heritability studies in the genus *Salix* have been conducted with *S. viminalis* (Rönnerberg-Wästljung *et al.*, 1994; Rönnerberg-Wästljung and Gullberg, 1999; Hallingbäck *et al.*, 2015) and *S. eriocephala* (Lin and Zsuffa, 1993; Cameron *et al.*, 2008). Trait heritability in these species was analyzed with F_1 populations ranging between 40-60 families with a small number of individuals. Trait estimates differed between sites where reports of clonal mean H^2 for stem height, stem number and stem diameter ranged from 0.05-0.31 (Lin and Zsuffa, 1993) and also Cameron *et al.* (2008) reporting values between $H^2 = 0.22$ -0.34 and $h^2 = 0.16$ -0.32. Values for

biomass growth traits were reported between $H^2 = 0.55-0.79$ and $h^2 = 0.04-0.42$ analyzed in the *S. viminalis* association mapping panel (Hallingbäck *et al.*, 2015). However, heritabilities are not fixed and thus vary based on the genetic architecture of the trait. Additionally, estimates between different studies are not directly comparable as they only pertain to the specific species, populations, and environments for which they are considered. The biomass yield of shrub willow is a quantitative trait varying greatly between genotypes and has been shown to be affected by significant genotype-by-environment interactions (Mosseler *et al.*, 2014; Fabio *et al.*, 2016). The most significant trait for developing new cultivars is high yield, in combination with disease and pest resistance. Yield can be selected directly or indirectly, therefore it is advantageous to quantify sources of variation contributing to and highly correlated with yield. Determining the heritability and the components associated with biomass production is vital for efficient and accurate crop improvement.

To date, there have been no reports of a quantitative genetic analysis or genetic mapping study for *S. purpurea*. The particular objective of this study was to determine the extent to which the genotypic and phenotypic variation of morphological, phenological, physiological, and wood composition traits were associated with allelic variation with a set of genome-wide SNPs and test different GWAS models to identify candidate genes related to key bioenergy traits of interest with the ultimate goal of developing tools for MAS.

MATERIALS AND METHODS

Germplasm, Genotyping, and Phenotyping

The association population was composed of 251 genotypes of *S. purpurea* mentioned in Chapters 2 and 3. Sex (male/female) was scored for 251 clonally propagated plants including

110 US accessions/cultivars and 141 European accessions (Table A2.1). The same experimental design was used as described in the Materials and Methods of Chapter 3. In brief, 20-cm cuttings of 110 genotypes were hand planted using in a common garden design at three experimental sites (Table A3.2), Cornell University's New York State Agricultural Experiment Station (NYSAES) in Geneva, NY; Cornell University's Lake Erie Research and Extension Lab (CLEREL) in Portland, NY; and the West Virginia University Agronomy Farm in Morgantown, WV, in a randomized complete block design with six replicates of four-plant plots at each location in single-row spacing with 1.82 m between rows and 0.40 m between plants within rows. At the end of the establishment year, all plants were coppiced and trials were measured using the inner two plants of each four-plant plot across all sites in 2013 and 2014 for 110 individuals, where 24 traits were evaluated for biomass, morphological, phenological, physiological and wood composition as described (Tables A3.3, 4.1, 4.2), and then harvested and weighed in early 2015. A subset of traits, SPAD and rust severity, were measured and evaluated in 2015 (Table 4.2). Rust was surveyed by assessing all the plants in each plot at two locations (Geneva, NY and Portland, NY) (Chapter 3 Materials and Methods).

DNA isolation and genotyping was according to the Materials and Methods of Chapter 2. Briefly, genomic DNA was extracted from young leaf and shoot tips, flash frozen in liquid nitrogen and extracted using the DNeasy Plant Mini Kit (Qiagen; Valencia, CA, USA) and a modified purification protocol with dichloromethane and isopropanol (Chaves *et al.*, 1995). The quality of DNA was checked by agarose gel electrophoresis and quantity was estimated using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific; Wilmington, DE, USA). Library and sequencing preparation was based on a 48-plex genotyping-by-sequencing (GBS) protocol according to Elshire *et al.* (2011) with the restriction enzyme *ApeKI*. The resulting libraries were

sequenced on the Illumina HiSeq 2000 (Illumina, Inc.; San Diego, CA, USA) platform at the Cornell University Biotechnology Resource Center (Ithaca, NY, USA). Marker discovery and filtering was performed with TASSEL v3.0 GBS Discovery Pipeline and a custom perl script (Bradbury *et al.*, 2007). Raw reads from FASTQ files were trimmed to 64 bp and were processed to create a set of unique sequence tags, where the minimum count that a tag must be present across all samples was set to five, which resulted in 4,550,690 unique tags. Marker genotypes were called through physical alignment to the 94006 reference genome ("*Salix purpurea* v1.0, DOE-JGI," 2015) using BWA (Li and Durbin, 2009), which included a 20th naïve pseudomolecule made up of unassembled scaffolds. SNPs were retained in individuals with a call rate of <90% (removed with >10% missing data) and filtered with a minor allele frequency (MAF) <0.05, and genotypes were also screened for a minimum proportion of 50% missing data which provided a set of 25,556 SNPs.

LD, Genetic Parameters, and Association Analysis

To evaluate the resolution expected during the GWAS analysis, LD was estimated by calculating the square value of the correlation coefficient (r^2) between all pairs of markers using TASSEL (Bradbury *et al.*, 2007) and Haploview software (Barrett *et al.*, 2005). Only marker loci with minor allele frequency values above 0.05 and having at least 90% successful calls among the sample set were included for LD analyses. Significant r^2 values with (LOD>2) were included and plotted against the physical distance (bp) between markers and a non-linear regression curve was fitted to describe the trend of LD decay.

Mixed linear models were used to analyze all phenotypic data implemented in SAS[®] version 9.4 with the PROC MIXED statement and with the *lmer* model within the *lme4* package

in R (Bates *et al.*, 2015). The following linear mixed models were used to obtain accurate trait estimates for each genotype:

$$Y_{ijkl} = \mu + Y_i + L_j + R_k(L_j) + G_l + G_l L_j + L_j Y_i + G_l Y_i + G_l Y_i L_j + \varepsilon_{ijkl} \quad (1)$$

$$Y_{jkl} = \mu + L_j + R_k(L_j) + G_l + G_l L_j + \varepsilon_{jkl} \quad (2)$$

$$Y_{kl} = \mu + R_k + G_l + \varepsilon_{kl} \quad (3)$$

where Y_{ijk} is the phenotypic trait measured for the i^{th} year, k^{th} replicate nested in the j^{th} location, G_l is the random effect of genotype l , $G_l L_j$ is the random interaction effect of the l^{th} genotype and the j^{th} location, the interaction of location by year ($L_j Y_i$) and genotype by year and $G_l Y_i$ with ε_{ijk} as the experimental error, and μ as the population mean (Model 1). For traits that were only measured during a single year, a reduced model was used (Model 2) and for traits only measured in a single location and year, Model 3 was used. Estimates for the variance components were obtained by restricted maximum likelihood (REML) and were used to estimate broad-sense heritability (H^2) for a given trait at a single location as,

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2/r} \quad (4)$$

where σ_g^2 and σ_ε^2 are the genotype and error variances, respectively, r is the number of replicates. For heritability estimates combining locations, heritability was calculated as,

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gl}^2/l + \sigma_\varepsilon^2/lr} \quad (5)$$

where σ_{gl}^2 is the genotype by environment interaction and l is the number of locations.

Markers were also used to calculate a normalized identity-by-state (IBS) kinship matrix in TASSEL 5.2.18 to estimate narrow-sense heritability using the *heritability* package in R

(Kruijer *et al.*, 2015),

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\varepsilon^2}$$

Best linear unbiased predictor (BLUP) estimates for all traits were used for GWAS performed using the GAPIT package in R (Lipka *et al.*, 2012; Tang *et al.*, 2016). In order to control confounding effects and improve statistical power while reducing the incidence of inflated *P*-values, three mixed linear models were used (Yang *et al.*, 2014): mixed linear model (MLM), compressed linear mixed model (CMLM), and settlement of MLM under progressively exclusive relationship (SUPER) (Wang *et al.*, 2014). The MLM model EMMAX (Kang *et al.*, 2010) was used to control for cryptic relatedness and population structure using an IBS matrix which has the equivalent statistical power of a standard MLM. A CMLM model was used which has been demonstrated to improve statistical power by 5 to 15% (Zhang *et al.*, 2010) by clustering individuals into groups based on their relationship using all markers by replacing individual genetic effects. A third model, SUPER (Wang *et al.*, 2014), was tested which uses only selected associated markers as quantitative trait nucleotides (QTNs) to derive kinship which improves statistical power at detecting marker-trait associations. To determine which models and corrected parameters best fit the data, observed and expected $-\log_{10}(p\text{-value})$ distributions for each SNP association were plotted against each other (quantile-quantile (QQ)-plots). Nominal and false discovery rate (FDR) adjusted *p*-values were considered using Bonferroni-corrected *p*-values at the 5% level of significance.

RESULTS

Heritability Estimates

Broad-sense heritability estimates (H^2) calculated for *S. purpurea* in the three association

trials were relatively consistent for a given trait, but with variability in year to year estimates (Table 4.1). Among most traits there were lower estimates of h^2 compared to H^2 , as expected, except for LFW at Portland in 2013 and Geneva in 2014 which also had high standard errors relative to the heritability estimate. H^2 values were greatest for HT (0.91) observed across years and sites and was greater overall for biomass/morphology traits than other categorized characters. Phenology H^2 estimates were moderately high (0.65-0.89) while h^2 values were low to moderate (0.32-0.55). Among the physiological traits, SPAD measurements varied greatly from year to year across sites, but remained fairly consistent between years within each site with Morgantown having greater values compared to the other two sites for these trait heritability values. For wood components, estimates of H^2 based on clonal means was moderate to high, while h^2 estimates were consistently lower, where values across years varied the most for LIG content and SPAD.

Table 4.1 Broad-sense (H^2) and marker based narrow-sense (h^2) heritability estimates of phenotypic traits from the *S. purpurea* association population (n=110) at three locations.

Trait ^a	Geneva		Portland		Morgantown	
	H^2	h^2	H^2	h^2	H^2	h^2
2013						
SDIA	0.79 (0.02)	0.41 (0.03)	0.77 (0.03)	0.37 (0.06)	0.71 (0.03)	0.35 (0.11)
SA	0.74 (0.16)	0.33 (0.02)	0.68 (0.12)	0.26 (0.05)	0.59 (0.10)	0.19 (0.09)
HT	0.91 (0.05)	0.61 (0.12)	0.84 (0.03)	0.48 (0.13)	0.25 (0.29)	0.09 (0.17)
IL	0.78 (0.16)	0.37 (0.21)	0.79 (0.05)	0.40 (0.09)	0.67 (0.06)	0.26 (0.13)
SNo	0.78 (0.09)	0.38 (0.04)	0.60 (0.02)	0.20 (0.05)	0.60 (0.01)	0.23 (0.12)
LFL	0.87 (0.11)	0.54 (0.14)	0.66 (0.12)	0.26 (0.06)	-	-
LFW	0.78 (0.06)	0.37 (0.01)	0.12 (0.15)	0.20 (0.23)	-	-
LFA	0.85 (0.03)	0.48 (0.02)	0.44 (0.07)	0.02 (0.11)	-	-
LFP	0.85 (0.16)	0.48 (0.03)	0.17 (0.23)	0.04 (0.12)	-	-
LFWT	0.68 (0.11)	0.48 (0.08)	0.73 (0.13)	0.49 (0.10)	-	-
SLA	0.19 (0.23)	0.05 (0.21)	0.17 (0.10)	0.03 (0.11)	-	-
AugSPAD	0.33 (0.15)	0.08 (0.23)	0.10 (0.19)	0.02 (0.16)	0.74 (0.15)	0.32 (0.18)
SeptSPAD	0.19 (0.19)	0.04 (0.25)	0.66 (0.12)	0.25 (0.07)	-	-
g_s	0.50 (0.16)	0.11 (0.13)	0.32 (0.12)	0.10 (0.21)	0.39 (0.23)	0.10 (0.25)
HEMI	0.55 (0.05)	0.23 (0.03)	0.51 (0.02)	0.20 (0.11)	0.54 (0.11)	0.22 (0.15)
CELL	0.41 (0.03)	0.15 (0.04)	0.40 (0.04)	0.14 (0.03)	0.42 (0.03)	0.16 (0.05)
LIG	0.39 (0.50)	0.14 (0.02)	0.46 (0.12)	0.18 (0.02)	0.48 (0.04)	0.19 (0.03)
ASH	0.62 (0.14)	0.29 (0.02)	0.50 (0.13)	0.20 (0.11)	0.54 (0.03)	0.24 (0.04)
SPGR	0.34 (0.04)	0.11 (0.03)	0.40 (0.06)	0.14 (0.07)	0.40 (0.02)	0.14 (0.04)
2014						
SDIA	0.76 (0.03)	0.38 (0.02)	0.83 (0.05)	0.46 (0.04)	0.76 (0.04)	0.41 (0.13)
SA	0.77 (0.11)	0.36 (0.12)	0.71 (0.05)	0.29 (0.13)	0.62 (0.02)	0.22 (0.05)
HT	0.91 (0.02)	0.63 (0.01)	0.91 (0.04)	0.63 (0.02)	0.82 (0.01)	0.43 (0.06)
IL	0.91 (0.23)	0.38 (0.06)	0.53 (0.05)	0.16 (0.15)	-	-
SNo	0.79 (0.15)	0.39 (0.05)	0.55 (0.11)	0.17 (0.02)	0.64 (0.12)	0.23 (0.13)
CDIA	0.76 (0.19)	0.34 (0.21)	0.23 (0.16)	0.06 (0.03)	0.50 (0.03)	0.14 (0.05)
CFOR	0.77 (0.17)	0.35 (0.16)	0.46 (0.25)	0.13 (0.10)	0.46 (0.01)	0.13 (0.06)
LFL	0.75 (0.21)	0.33 (0.12)	0.76 (0.13)	0.34 (0.06)	0.61 (0.05)	0.20 (0.11)
LFW	0.30 (0.23)	0.30 (0.25)	0.09 (0.15)	0.01 (0.05)	0.49 (0.05)	0.14 (0.17)
LFA	0.75 (0.06)	0.33 (0.03)	0.79 (0.09)	0.39 (0.05)	0.56 (0.08)	0.18 (0.10)
LFP	0.71 (0.18)	0.28 (0.03)	0.42 (0.13)	0.11 (0.13)	0.37 (0.19)	0.09 (0.03)
LFWT	0.83 (0.15)	0.49 (0.09)	0.82 (0.15)	0.50 (0.09)	0.72 (0.14)	0.49 (0.15)
SLA	0.36 (0.22)	0.09 (0.31)	0.49 (0.19)	0.13 (0.19)	0.26 (0.21)	0.05 (0.06)
YLD	0.84 (0.15)	0.47 (0.16)	0.87 (0.10)	0.55 (0.12)	0.64 (0.09)	0.24 (0.12)
VPH	0.88 (0.11)	0.55 (0.05)	-	-	0.74 (0.16)	0.32 (0.23)
FPH	0.79 (0.16)	0.38 (0.06)	-	-	0.76 (0.13)	0.35 (0.13)
AugSPAD	0.79 (0.23)	0.39 (0.19)	0.69 (0.23)	0.28 (0.15)	-	-
SeptSPAD	0.51 (0.13)	0.15 (0.09)	0.59 (0.25)	0.19 (0.08)	-	-
g_s	0.58 (0.12)	0.17 (0.11)	0.49 (0.09)	0.13 (0.10)	0.29 (0.13)	0.04 (0.28)

Table 4.1 (Continued)

HEMI	0.70 (0.05)	0.36 (0.06)	0.62 (0.04)	0.28 (0.06)	0.58 (0.05)	0.26 (0.13)
CELL	0.69 (0.06)	0.37 (0.11)	0.50 (0.05)	0.20 (0.09)	0.59 (0.06)	0.27 (0.14)
LIG	0.70 (0.10)	0.37 (0.06)	0.67 (0.12)	0.34 (0.05)	0.63 (0.13)	0.30 (0.16)
ASH	0.71 (0.13)	0.38 (0.10)	0.51 (0.08)	0.21 (0.10)	0.54 (0.11)	0.23 (0.18)
SPGR	0.87 (0.13)	0.63 (0.11)	0.64 (0.07)	0.30 (0.03)	0.58 (0.05)	0.25 (0.06)
2015						
RUST	0.79 (0.11)	0.68 (0.15)	0.81 (0.09)	0.77 (0.19)	-	-
SPAD	0.59 (0.23)	0.49 (0.13)	0.57 (0.25)	0.45 (0.13)	0.58 (0.35)	0.35 (0.29)

^aPhenotypic traits were measured in years 2013, 2014, and 2015. See Chapter 3 Materials and Methods for trait abbreviations and definitions.

H^2 , broad-sense heritability estimate

h^2 , marker-based narrow-sense heritability estimate using genotypic means

Estimation of standard errors are given in parentheses

Trait Correlations

Significant positive and negative correlations were detected between all traits, where morphology and wood composition had the strongest positive correlation with YLD. Comparisons between years and locations for each trait, revealed significant positive relationships of SDIA, SA, HT, SNo, LFL, LFA, and HEMI and CELL with YLD. Pairwise comparisons between these traits also showed positive associations. Leaf area and LFP were positively correlated with AugSPAD measurements, stem HT, IL and YLD, but were very low to moderate in magnitude. Using mean values, the most significant positive correlations with YLD came from SA ($R^2=0.73$) and HT ($R^2=0.68$). This trend was also consistent across years for each trait. Significant negative correlations of traits with YLD included LIG, ASH, and SPGR with an increasingly negative correlation from the first to the second year. Hemicellulose and CELL were always significantly positively associated with each other as was the relationship between LIG and ASH. However, a significant negative correlation of HEMI and CELL versus LIG and ASH was always observed within and across years and locations.

GBS Analysis and Marker Distribution

The dataset obtained using the reference based GBS filtering pipeline yielded 25,566 markers with an average nucleotide diversity of $\pi=0.30$ (Figure 4.1). Filtering criteria were selected to remove markers with $MAF<0.05$, and the MAF distribution of the remaining markers at $MAF<0.10$ was 18.4% and the $MAF>0.25$ was 37.4% with the average $MAF=0.22$. The heterozygosity rates for each genotype ranged from 0.19 to 0.47 with an average heterozygosity of 0.29. The average density of SNPs corresponded to 1 marker every 13.6 kb. Linkage disequilibrium analysis showed that 32.9% of the marker pairs were in significant LD with a majority of markers exhibiting average r^2 values of 0.43. The GBS data in this study do not provide the ideal dataset to estimate LD due to small sample size and some unknown physical distances between markers due to 11% of gaps present in the physical assembly of the reference genome ("*Salix purpurea* v1.0, DOE-JGI," 2015). However, significant LD with $r^2\geq 0.2$ extended up to ~1kb (Figure A4.1).

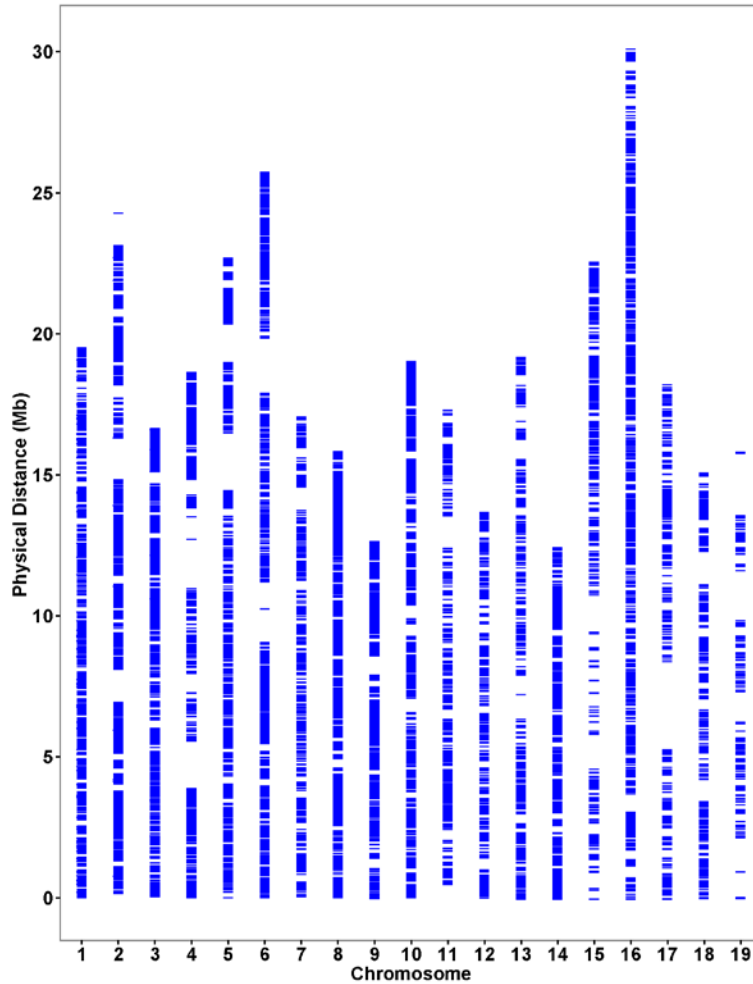


Figure 4.1 Distribution map of 25,566 GBS SNP markers across 19 *S. purpurea* chromosome-scale pseudomolecules

GWAS

Due to the small sample size of this population of $n=110$ for a suite of 24 traits and $n=251$ for sex determination, GWAS analyses had very limited power to detect significant associations of small or moderate effect, as expected (Figure A4.2). In order to avoid detection of false positives, a series of MLMs were used to increase statistical power and correct for kinship (Figure A4.3) and population structure. All models for each trait were evaluated based on the fit of the model to the data and resulted in the SUPER method performing well overall with an

average genomic inflation factor of $\lambda_{gc}=1.13$ (Figure A4.4-A4.5). Overall, 95 significant associations ($p<10E-5$) were detected on 17 of the 19 *Salix* chromosomes, with no associations mapping to chromosomes 10 and 16. Significant associations were also detected on the 20th naïve pseudochromosome made up of unassembled scaffolds. Fifty-seven of these markers were associated with eight traits (Table 4.2, Figure A4.5), where three of these SNPs reached genome-wide significance after Bonferroni correction at $\alpha=0.05$ and 51 SNPs had an estimated FDR<0.05. The remaining 38 SNPs corresponded to six genomic locations associated with plant sex on chromosomes 1, 2, 3, 9, 13, 15, 16, and on the 20th pseudochromosome (Figure 4.2). Twenty-three unique candidate genes were associated with these markers (Table 4.3), but 10 of the markers were unable to be matched to any candidate genes. The majority of these SNPs that were associated with the sex phenotype clustered within a 15.9 Mb region on chromosome 15 and were the strongest significant associations found in the study.

Table 4.2 Statistically significant marker-trait associations for 110 genotypes and candidate genes.

Trait	Chromosome	Position	P-value	Allele	MAF	Description
CDIA	1	17,122,175	1.75E-08**	C/T	0.05	Loricrin-like protein,
CDIA	2	8,133,354	7.48E-11**	T/C	0.09	plant/MNA5-17 protein
CDIA	2	19,923,625	4.30E-08**	T/C	0.05	Organic anion transporter,
CDIA	3	6,972,881	1.43E-07**	T/C	0.09	light harvesting-like protein
CDIA	4	18,590,829	1.28E-07**	A/G	0.05	NA
CDIA	6	1,278,153	1.28E-07**	C/T	0.05	HAT family dimerization protein
CDIA	11	8,157,827	2.06E-07**	T/C	0.07	Transmembrane protein,
CDIA	12	9,793,748	2.40E-07**	T/G	0.05	Membrane steroid-binding protein
CDIA	14	11,256,441	3.78E-07**	G/A	0.05	Nucleoside triphosphate pyrophosphohydrolase
CDIA	16	22,419,773	2.73E-07**	G/C	0.05	Glycoside hydrolase family 9 protein
CDIA	18	12,673,620	4.82E-09**	G/C	0.05	Lysosomal pro-X carboxypeptidase
LFL	1	2,925,313	1.99E-07**	T/C	0.14	RNase L inhibitor ABC domain protein
LFL	1	8,804,516	7.12E-08**	G/C	0.24	RNase L inhibitor ABC domain protein
LFL	1	8,806,048	7.12E-08**	T/G	0.30	RNase L inhibitor ABC domain protein
LFL	1	16,092,322	2.37E-07**	C/T	0.05	MYB transcription factor-like protein,
LFL	2	4,357,552	3.37E-08**	C/A	0.08	RNase L inhibitor ABC domain protein
LFL	3	166,519	2.51E-08**	G/C	0.13	kinesin motor domain protein
LFL	3	4,041,449	1.41E-07**	G/T	0.05	Reverse transcriptase-like protein
LFL	4	905,677	3.37E-08**	A/G	0.22	LRR receptor-like kinase
LFL	4	3,285,295	3.37E-08**	A/G	0.10	Plant UBX domain protein
LFL	5	11,233,191	3.37E-08**	T/A	0.25	Heat shock protein 70 (HSP70)-interacting protein,
LFL	6	20,993,081	2.06E-07**	G/A	0.05	Transcription factor ORG2,
LFL	6	21,121,321	2.51E-08**	C/T	0.20	Transcription factor ORG2,
LFL	7	4,090,929	3.37E-08**	A/C	0.23	Serine/Threonine kinase domain protein
LFL	7	16,666,875	2.51E-08**	G/A	0.07	Type I inositol 1,4,5-trisphosphate 5- phosphatase
LFL	8	1,930,707	2.67E-08**	G/A	0.06	Ras small GTPase family Ras protein
LFL	8	8,781,244	1.12E-07**	T/C	0.05	BRO1-like domain

Table 4.2 (Continued)

LFL	8	14,848,119	3.92E-08**	C/T	0.14	Phosphorylase superfamily protein
LFL	8	15,669,429	2.54E-07**	A/G	0.05	Cyclin-dependent kinase inhibitor siamese protein,
LFL	9	330,288	3.37E-08**	T/A	0.08	Cytochrome P450 family protein
LFL	9	5,080,779	2.51E-08**	G/T	0.07	Thylakoid lumenal 19 kDa protein
LFL	11	16,794,098	7.69E-08**	A/T	0.05	Peroxidase
LFL	13	3,481,405	2.98E-08**	G/C	0.27	Hypothetical protein
LFL	13	17,797,130	3.37E-08**	A/T	0.25	DEAD-box ATP-dependent RNA helicase
LFL	14	4,789,408	3.37E-08**	G/A	0.07	ATP-dependent DNA helicase
LFL	14	6,127,386	2.11E-07**	T/G	0.05	Glutathione peroxidase
LFL	15	20,421,631	3.37E-08**	A/G	0.29	Calnexin
LFL	17	16,991,024	2.51E-08**	A/G	0.11	NA
LFL	18	4,275,430	3.37E-08**	T/A	0.09	Lactoylglutathione lyase
LFL	20	2,589,718	3.37E-08**	T/C	0.29	NA
LFL	20	33,305,917	4.47E-08**	T/C	0.10	NA
LFL	20	39,410,449	3.36E-07**	C/G	0.05	NA
LFWT	8	4,155,140	4.40E-08**	C/T	0.21	Hypothetical protein
RUST	1	4,867,988	9.50E-07*	T/G	0.50	DNA replication licensing factor/MCM complex/DNA helicase
RUST	2	9,632,552	1.14E-05*	A/T	0.50	Tetratricopeptide repeat protein
RUST	2	11,458,061	1.34E-06*	T/C	0.50	NA
RUST	2	11,784,305	5.76E-06*	G/T	0.50	NA
RUST	2	12,498,525	2.10E-06*	C/T	0.50	Xylem serine proteinase/Subtilisin-like protease
RUST	2	12,748,684	1.57E-06*	G/A	0.50	Transmembrane protein,
SDIA	3	8,622,959	2.75E-07**	A/C	0.41	Reverse transcriptase-like protein
SeptSPAD	2	19,326,675	2.62E-07**	G/T	0.32	DNAJ heat shock amino-terminal domain protein,
SeptSPAD	8	12,705,219	2.77E-07**	A/G	0.30	RING zinc finger protein,
VPH	15	5,890,467	7.01E-08**	G/A	0.26	ATP-binding protein,

Table 4.2 (Continued)

VPH	20	9,403,446	2.53E-07**	G/T	0.12	glucan endo-1,3-beta-glucosidase
YLD	3	8,620,442	3.76E-07**	T/C	0.41	BZIP transcription factor
YLD	3	8,622,898	4.03E-08**	A/G	0.42	BZIP transcription factor
YLD	9	11,500,126	6.32E-08**	A/C	0.41	NA

Most of the marker associations that reached genome-wide significance were located within genic regions (5'UTR, CDS, intron, or 3'UTR') and candidate genes were inferred based on functional annotation of the *S. purpurea* genome. Ten candidate genes were identified for crown diameter, four of which were only putative proteins. Thirty one SNPs were significantly associated with LFL and matched 24 unique candidate genes, which were mainly related to several classes of receptor kinases. A significant marker was also found be associated with LFWT, but only matched a hypothetical protein with no known similarities using a BLAST search. Six markers were significantly associated with RUST that matched four candidate genes, potentially linked to a tetratricopeptide repeat protein (Figure 4.3). Stem diameter was found to be associated with a reverse transcriptase-like protein and SeptSPAD and VPH each had two significant markers associated with candidate genes involved with four different classes of proteins. Also, three significant markers were associated with biomass yield with two of the markers located on chromosome 3 separated by ~2.5 kb that were functionally annotated as BZIP transcription factors. The third marker associated with yield was located on chromosome 9 but did not fall within a genic region of the *S. purpurea* genome.

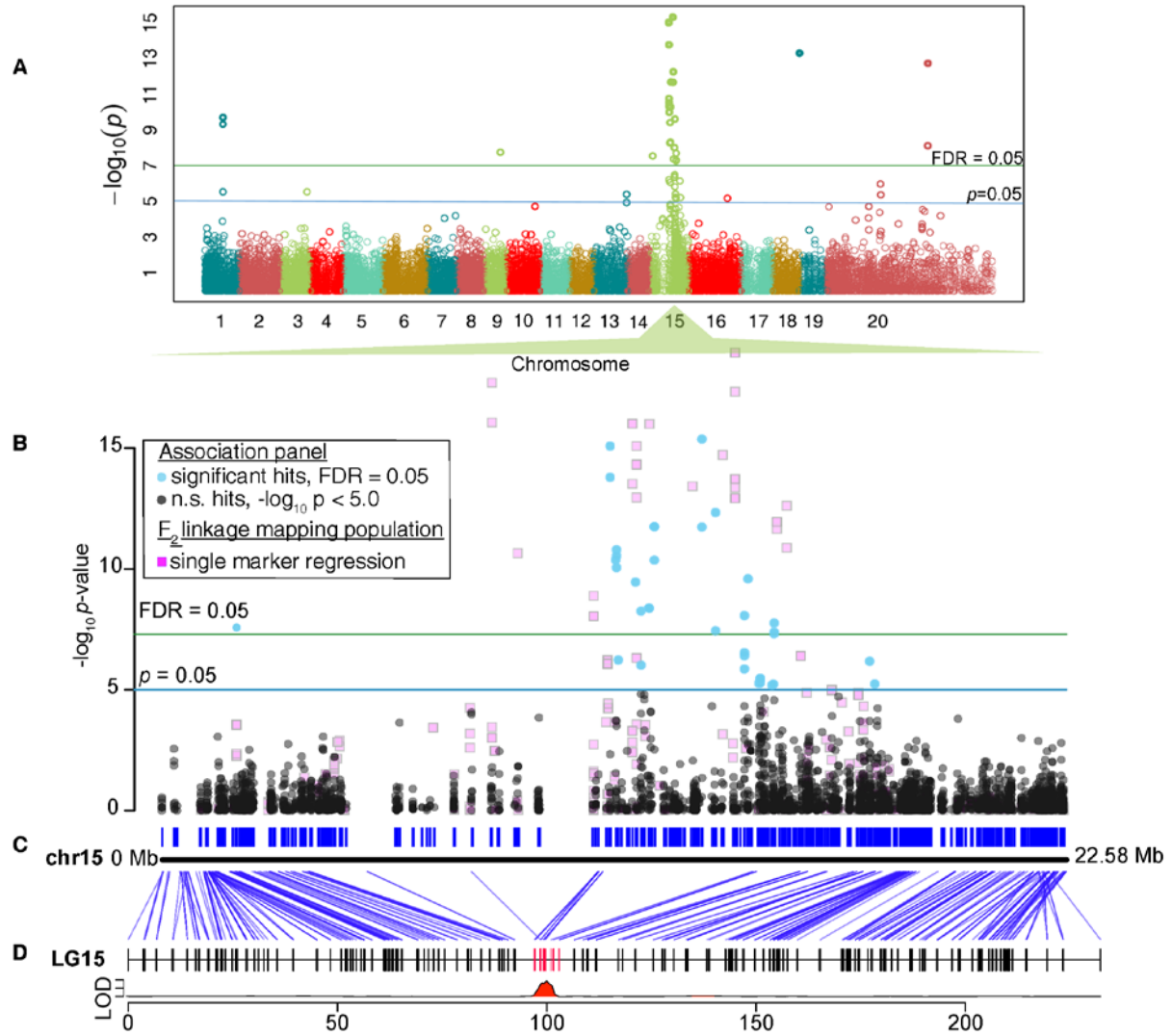


Figure 4.2 A) Genome-wide association results for sex phenotypes of 251 natural accessions with 25,566 SNPs across 19 *S. purpurea* chromosomes and a 20th naïve pseudochromosome represented by concatenated unassembled scaffolds. B) Significant GWAS SNP associations shown in blue circles on chromosome 15 for unadjusted p -values < 0.05 and at a false discovery rate (FDR) significance threshold of 5%. Red squares show significant QTL markers from an F_2 full-sib linkage mapping population ($n=497$) associated with the sex phenotype on chromosome 15. Alignment of C) the physical distribution (Mb) of SNPs along chromosome 15 with the D) genetic distance (cM) on LG15.

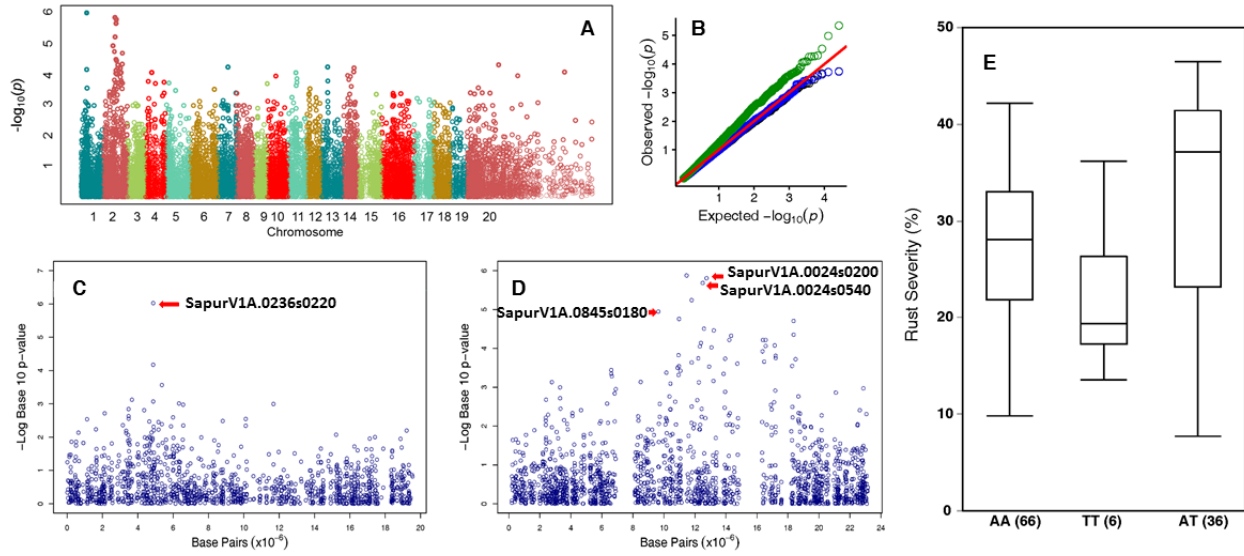


Figure 4.3 A) Manhattan plot of associations of 25,566 SNPs for rust severity based on the SUPER model with $\lambda=1.12$. The 20th naïve pseudo-chromosome represents concatenated unassembled scaffolds. B) QQ-plot for three GWAS MLM models for rust severity (black circles correspond to MLM, blue circles for CMLM, and green circles for SUPER model. Significant ($p<0.05$) annotated SNPs are shown on C) chromosome 1 and D) chromosome 2. E) Boxplots show raw effect size for SNP found on chromosome 2 within the gene SapurV1A.0845s0180 with each allelic variant (x-axis) and the sample size (n) within the population shown in brackets

Table 4.3 Statistically significant marker-trait associations for 251 genotypes and candidate genes for plant sex determination.

Chromosome	Position (bp)	P-value	Allele	MAF	Description
1	10,104,254	4.79E-10	C/T	0.23	NA
3	14,521,607	2.81E-06	T/C	0.11	Extra response regulator/ histidine kinase.
9	8,880,998	1.70E-08	A/G	0.39	NA
13	18,626,508	3.88E-06	T/C	0.28	Multicopper oxidase/laccase
15	1,878,562	2.73E-08	G/A	0.25	ATP-dependent RNA helicase
15	4,619,319	5.14E-06	A/G	0.47	NA
15	8,242,679	1.25E-07	T/C	0.06	NA
15	8,257,798	2.58E-14	A/G	0.42	Carbohydrate-binding module family protein
15	8,276,012	1.95E-13	A/G	0.25	Carbohydrate-binding module family protein
15	11,228,738	9.27E-16	T/C	0.33	NA
15	11,372,200	4.46E-11	A/G	0.19	Heat shock protein/ ATP-dependent clp protein
15	11,389,975	5.62E-13	G/T	0.24	Replication protein A/DNA-binding subunit
15	11,433,697	5.93E-07	G/T	0.35	NA
15	11,862,825	3.88E-10	A/C	0.46	Serine/Threonine kinase domain protein
15	12,004,173	3.97E-12	G/A	0.23	Polyprenyl synthetase
15	12,206,456	4.69E-09	C/T	0.33	Myb transcription factor,
15	12,335,923	2.07E-12	A/T	0.20	DNA-binding protein,
15	13,866,709	4.04E-08	A/T	0.13	High mobility group family protein
15	14,034,362	1.53E-06	C/T	0.33	NA
15	14,589,419	9.50E-09	G/T	0.20	NA
15	14,680,425	2.33E-10	G/T	0.14	Laccase
15	14,958,469	3.18E-06	T/G	0.04	Receptor-like kinase,
15	14,960,487	2.35E-10	C/T	0.18	Receptor-like kinase,
15	14,987,112	3.61E-06	C/T	0.05	Hypothetical protein
15	15,282,435	6.96E-06	A/G	0.13	Hypothetical protein
15	15,317,274	6.23E-06	C/T	0.14	Geranylgeranyl diphosphate reductase
15	15,334,002	6.75E-13	T/C	0.24	Phytochrome kinase substrate protein,
15	16,203,979	4.04E-07	T/G	0.23	HIPL1 protein,
15	16,383,583	3.03E-06	C/T	0.47	Hypothetical protein
15	16,470,762	1.19E-08	A/G	0.20	Phosphatidylinositol 3-kinase

Table 4.3 (Continued)

15	16,787,477	4.89E-07	T/C	0.42	BAH and TFIIS domain protein
15	17,854,339	5.95E-06	A/G	0.08	ETEA-like protein,
16	22,215,470	6.45E-06	C/T	0.25	Magnesium transporter
19	69,904	4.90E-14	A/G	0.22	Transmembrane protein,
20	30,903,742	4.12E-06	A/G	0.34	Transmembrane protein,
20	30,909,933	1.03E-06	A/C	0.19	DUF789 family protein
20	58,058,949	1.78E-13	C/T	0.35	NA
20	58,058,965	7.39E-09	T/G	0.48	NA

DISCUSSION

In this study a natural population of *S. purpurea* was used to estimate genetic parameters and conduct a GWAS to identify candidate genes and uncover the role of genetic variants for complex traits related to biomass production.

Overall H^2 estimates were moderate with h^2 marker-based estimates showing low to moderate values. The estimates of genetic variation in this study were highly significant for most traits and suggest that direct phenotypic selection for these traits should result in significant gains. These values imply that at least some of the genetic variation in these traits is due to common genetic polymorphisms. The difference between h^2 and H^2 , similar to the "missing heritability" in studies with complex traits such as human height (Yang *et al.*, 2010) likely stems from non-additive factors (e.g. epistatic interactions) as well as rarer causal polymorphism not tagged by the common GBS markers used in this study. If better estimates are to be obtained, larger sample sizes and greater marker density are needed to disentangle these explanations.

It was observed that some of the H^2 estimates were higher than expected or what has been previously reported for similar traits (Tsarouhas *et al.*, 2002; Cameron *et al.*, 2008; Ghelardini *et al.*, 2014). This may be due to dominance or epistatic variance contributing to the genetic architecture of these traits. The biggest factor for this dataset using GBS markers would be the fact that only very common alleles were used to estimate the kinship matrix. This dictates that anything causal under 5% MAF will be uncaptured, and is probably a substantial portion of variation. The h^2 and H^2 estimates were similar, but with substantial standard errors relative to the estimates given the small sample size of the population, the differences are difficult to interpret. Additionally, genotyping error will also likely degrade h^2 further, especially if it is non-random (i.e., shared errors disrupting the relatedness matrix). That is not to say that

dominance and epistasis are not likely to be important, but instead it is difficult to conclude this when comparing a purely phenotypic estimate to an estimate that relies on markers plus phenotypes.

Previous quantitative genetics studies of multiple F₁ full-sib populations of *S. viminalis* have shown a smaller proportion of genetic variance relative to environmental variance, accounting for up to 78% of the total variation when analyzing biomass characteristics across two sites (Cameron *et al.*, 2008). Growth phenology has been shown to be under relatively high genetic control, but still containing a significant proportion of variation explained by genotype x environment interaction (Ghelardini *et al.*, 2014). In the current study, the overall moderate broad-sense heritability estimates on clonal means suggests that environmental factors are influencing these traits and that multiple replications in several environments should be used for screening and selection since variances and heritability estimates are specific to the populations and environments being measured.

Phenology measurements are important traits as they provide an indication of determining the length of the growing season. Traits measured in this study exhibited relatively high heritability and genetic variance for vegetative and floral phenology. It has been shown that light use efficiency is an important factor contributing to aboveground biomass and that early bud break may contribute to this production (Tharakan *et al.*, 2008). However, it has been suggested that other phenological events (i.e. leaf unfolding and duration, growth cessation and leaf abscission) affect annual biomass production, where growth cessation and late season leaf retention may also impact aboveground growth as well as nutrient recycling and storage (Weih, 2009). These traits were not measured in this study and may partially explain statistically weak or negative correlations with yield.

There were significant effects of location by year variance for all physiological traits where it appears differences in growth conditions of a particular year varied more than the experimental sites. These physiological measurements had relatively low genetic variances accounting for only 3-8% for the three traits in this category. The growing season of 2014 was hot and humid and provided ideal weather conditions which may be contributing to this highly significant effect. Stomatal conductance was measured to investigate variability in water loss and CO₂ uptake which showed low to moderate heritability and only had weak positive correlations with several other traits. However, in order to observe well adapted genotypes under environmental stress and understand the mechanistic or deterministic basis, more detailed treatment-control experiments are needed with emphasis on drought conditions. Drought response is a characteristic that is important to the breeding of shrub willow since some of the targeted environments for plantings are on marginal lands (Stolarski *et al.*, 2011; Amichev *et al.*, 2012) which can be too dry because of low gravimetric soil moisture. Water use efficiency is likely a quantitative trait that could be influenced by many factors such as root distribution, stomatal conductance and photosynthetic rate during limited water availability. The ability to conserve water under these circumstances and environments is a desirable trait that requires further study.

For association mapping, a large number of GBS markers were generated using a single-methylation sensitive restriction enzyme, however, the density of markers was relatively low given the size of the *S. purpurea* reference genome. This may affect the resolution of mapping and limit the ability for the complete dissection of complex trait architecture, especially if polymorphisms of small effects are located outside of genes, but may be improved upon with the use of double-digestion with two restriction enzymes. However, a number of significant genome-

wide SNPs were detected despite the small sample size of the population. Multiple significant associations were found which may be due to longer ranges of LD on certain chromosomes that were not detected, where LD between distant loci can inflate single locus test statistics (Thomas *et al.*, 2011). Low LD was observed in this study, but many highly heterozygous, outcrossing plants such as tree species display rapid LD decay which was reported around 2.6 kb for *P. deltoides*, and greater than 1 kb in *P. trichocarpa* (Slavov *et al.*, 2012), compared to self-pollinated species such as rice (75-150 kb) (Huang *et al.*, 2010) and maize (1.50-10 kb) (Yan *et al.*, 2009). Despite this complexity, this suggests that alternative SNPs may be located in the proximity to a significantly associated SNP even though that marker itself might not be causative. It should be noted that some of the genetic architectures of the traits in this study are highly complex, where tens or hundreds of causative polymorphisms with minor effects may exist genome-wide.

Among the associations observed in this study, promising observations were found for CDIA, RUST, and sex. A significant SNP was found on chromosome 16 associated with CDIA and was located in the coding sequence for a glycoside hydrolase gene, specifically GH9 which is the second largest cellulose family (Davies and Henrissat, 1995). Studies for this enzyme family indicate that they are involved in cell wall modification during fruit softening, abscission, growth, and wood formation (Urbanowicz *et al.*, 2007; Du *et al.*, 2015). Protein homologs for this gene matched other known GH9 proteins in *Populus trichocarpa* and *Theobroma cacao*. Based on sequence alignments and Pfam database searches, the candidate gene found in *S. purpurea* belongs to subclass B, which comprises secreted proteins with only one catalytic domain (Urbanowicz *et al.*, 2007). GH9B, synonymous with endo-1,4- β -glucanase 11, has been reported with activities for cello-oligosaccharide release and xyloglucan cleavage in

plants, but the GH9 superfamily has yet to be characterized in *Salix*. One well-studied GH9 clade includes a membrane-associated endoglucanase (KORRIGAN) that is part of the cellulose synthesis complex and that influences the organization of cellulose in the wall. This candidate gene may have relevant functional information for CDIA since shrub willow contains numerous woody stems that can have various sweeping branching angles from the base of the crown which would require growing the cell wall to withstand the high tensile forces generated by cell wall stress for relaxation and cell wall expansion. The results in Chapter 2 revealed consistent trends of significant male bias for greater CDIA and subsequent shallower branching angle. This gene may have future implications on targeting mechanical properties of wood to influence biomechanical strength and therefore directly influencing plant architecture.

Another significant SNP was associated with RUST severity on chromosome 2 that fell within the coding region of the gene for a tetratricopeptide repeat protein (TPR). The TPR motif is a 34 amino acid consensus sequence that has been well characterized, with functions serving to recognize pathogen infection and trigger plant autoimmune response (Schapire *et al.*, 2006) which has been shown to confer resistance to rust caused by *Melampsora lini* in flax (Lawrence *et al.*, 2010). It has been more recently shown that a mutation of the gene encoding a TPR domain-containing protein, SRFR1 resulted in autoimmune responses owing to transcriptional upregulation of several co-regulated R genes (Kwon *et al.*, 2009; Kim *et al.*, 2010). In this study, it was seen that there was a dramatic difference of allelic effect on rust severity (Figure 4.3), where genotypes homozygous for either the A or T allele had less incidence of rust compared to those that had the heterozygous allele. Additionally, the marker appeared to be a non-synonymous SNP, changing glutamic acid to valine, which will require further study since this marker did not coincide with any previously reported rust resistant loci for willow (Hanley and

Karp, 2013).

Lastly, a strong association with plant sex was observed with a significant association peak on chromosome 15. This has strong implications for basic biological understanding of plant sex evolution and development. While there were no specific candidate genes that have known functions related to sex determination in *Salix*, the strong association for the locus also coincides with all previous reports of being located in chromosome 15 across multiple species (Hou *et al.*, 2015; Pucholt *et al.*, 2015; Chen *et al.*, 2016). Additionally, a comparison of QTLs mapped for sex from an F₂ full-sib population (Chapter 2) also confirmed the location of the sex chromosome and locus due to significant overlapping of markers from both mapping populations.

CONCLUSION

In summary, high quality phenotypic data for 24 traits in a population of 110 *S. purpurea* individuals were associated with GBS to estimate genetic parameters and identify candidate genes for eventual use for MAS. However, because of the small population size, the power of detection was low, but promising candidate genes from GWAS suggest this approach will be valuable upon further expansion of the population and density marker coverage.

CHAPTER 5 - Future Directions

The work presented here reveals the diversity of *Salix* through characterization of the population structure, evidence of complex interactions between sexual dimorphism and the genetics of sex determination, and the feasibility of implementing large-scale genomic mapping studies for trait discovery. These studies have prompted questions on how to integrate population genomics, association genetics, and comparative genomics to advance breeding efforts.

The long-term goal of this research is to enable development of affordable advanced molecular breeding methods for perennial bioenergy crops with superior performance, based on the understanding of the genetic regulation and physiological mechanisms controlling traits of interest. To achieve the full genetic potential in breeding second generation lignocellulosic crops, it is critical to utilize and integrate all resources. *Salix* lends itself as a model crop for studying population genetics, sexual dimorphism, and trait variability. The abundant genetic diversity, dioecy, and broad geographical distribution provide numerous outlets to apply this knowledge in improving crop productivity.

In my population of natural and native accessions, I observed that substantial population structure and differentiation exists and found evidence of subpopulations. Due to life history traits that promote the maintenance of genetic variation, including outcrossing mating systems, long generation times and extensive gene flow over large geographical distances, these characteristics are associated with high levels of genetic diversity at local and regional scales. The samples that were used in this study were collected from broadly geographically separated areas across many sites, but there was still clonality or close relatedness detected. Future efforts should focus on collection expeditions to sample the wider geographic distribution of the native and naturalized ranges with the goal of achieving at least 500 new unrelated accessions, which is

typically seen in population and association analyses of *Populus*, where larger populations provide greater statistical power to detect marker-trait associations.

The availability of whole-genome resequencing datasets is yielding important insight into a number of basic biological questions, including how natural selection shapes genome-wide patterns of variation and the genetic basis of climate adaptation. Efforts should be made to resequence the current association population with future re-sequencing efforts focused on new collections. In theory, this should reveal numerous genes involved in controlling natural variation for a number of ecologically relevant traits, such as bud break, bud set, leaf anatomy, and photosynthetic rate. Additionally, sequence variants other than SNPs have not been extensively documented and studied thus far in *Salix*, although evidence from other systems suggests that they can have large effects on phenotypic variation (Montgomery *et al.*, 2013). Identification of structural variation, such as presence/absence variants (PAV) and copy number variants (CNV) across individuals has led to the concept of the pan-genome of a species, which encompasses all genetic variation that can be found within a species. The pan-genome can further be subdivided into a core part, containing genomic regions present in all individuals, and a dispensable part, which contains the remaining genomic regions that are variably present in individuals. Given the high levels of genetic variation maintained in many *Salix* spp., it is likely single nucleotide variants (SNVs) are abundant and involved in mediating functionally important variation that can be detected using various genotyping-by-sequencing (GBS) methods, or methods such as exome-capture or sequence-capture (Zhou and Holliday, 2012).

These are all promising qualities for association genetics within the current population, enabling identification of informative candidate genes for future molecular breeding efforts for improved biomass yield. Although the strong effects of the candidates identified were substantial

and highly significant indicating usefulness in MAS, further validation is still warranted as they are based on very few accessions.

Additional research objectives related to these discoveries might include further investigation into the evolutionary forces that drove the great amount of genetic and phenotypic variation seen across many species and how this relates to the differences seen between sexes. More practical research objectives might include developing simple presence/absence markers for sex determination as a fast and efficient method for screening large breeding populations. This process could utilize GBS tags that are only present in males or females and then be used for further genotyping by using the tags for AmpSeq genotyping. In a broader context, additional, research objectives would relate to wood composition. Cellulose is the major wall polysaccharide in plants and has a wide application for biofuel, paper, and other chemical products. Due to their crystalline property, cellulose microfibrils are highly recalcitrant to biomass saccharification. Hence, understanding cellulose biosynthesis and crystallization is essential.

REFERENCES

- Åhman, I. (1997). Growth, herbivory and disease in relation to gender in *Salix viminalis* L. *Oecologia*, 111, 61-68.
- Ainsworth, C. (2000). Boys and girls come out to play: The molecular biology of dioecious plants. *Ann Bot*, 86, 211-221.
- Akagi, T., Henry, I.M., Tao, R., & Comai, L. (2014). A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science*, 346, 646-650.
- Allwright, M.R. and Taylor, G. (2016). Molecular breeding for improved second generation bioenergy crops. *Trends in Plant Science*, 21, 43-54.
- Alsos, I.G., Alm, T., Normand, S., & Brochmann, C. (2009). Past and future range shifts and loss of diversity in dwarf willow (*Salix herbacea* L.) inferred from genetics, fossils and modelling. *Global Ecol Biogeogr*, 18, 223-239.
- Alström-Rapaport, C., Lascoux, M., & Gullberg, U. (1997). Sex determination and sex ratio in the dioecious shrub *Salix viminalis* L. *Theor Appl Genet*, 94, 493-497.
- Amichev, B., Kurz, W., Smyth, C., & Van Rees, K. (2012). The carbon implications of large-scale afforestation of agriculturally marginal land with short-rotation willow in Saskatchewan. *GCB Bioenergy*, 4, 70-87.
- Angela, K., Steve, J.H., Sviatlana, O.T., William, M., Ming, P., & Ian, S. (2011). Genetic Improvement of Willow for Bioenergy and Biofuels. *J Integr Plant Biol*, 53.
- Aravanopoulos, F.A. and Zsuffa, L. (1998). Heterozygosity and biomass production in *Salix eriocephala*. *Heredity*, 81, 396-403.
- Argus, G.W. (1974). An experimental study of hybridization and pollination in *Salix* (willow). *Can J Bot*, 52, 1613-1619.
- Argus, G.W. (1997). Infrageneric classification of *Salix* (Salicaceae) in the New World. *Syst Bot Monogr*, 52, 1-121.

- Argus, G.W. (2007). *Salix* (Salicaceae) distribution maps and a synopsis of their classification in North America, north of Mexico. *Harvard Pap Bot*, 12, 335-368.
- Argus, G.W., Eckenwalder, J.E., & Kiger, R.W. (2010). Salicaceae. In Flora of North America Editorial Committee (Ed.), (Vol. 7). New York: Oxford University Press.
- Arnold, S.J. (1994). Bateman's principles and the measurement of sexual selection in plants and animals. *Am Nat*, 144, S126-S149.
- Arnoult, S. and Brancourt-Hulmel, M. (2015). A review on miscanthus biomass production and composition for bioenergy use: Genotypic and environmental variability and implications for breeding. *Bioenerg Res*, 8, 502-526.
- Bañuelos, M.-J., Sierra, M., & Obeso, J.-R. (2004). Sex, secondary compounds and asymmetry. Effects on plant–herbivore interaction in a dioecious shrub. *Acta Oecologica*, 25, 151-157.
- Barrett, J.C., Fry, B., Maller, J., & Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21, 263-265.
- Barrett, S.C.H., Yakimowski, S.B., Field, D.L., & Pickup, M. (2010). Ecological genetics of sex ratios in plant populations. *Philos T R Soc B*, 365, 2549-2557.
- Bassi, F.M., Bentley, A.R., Charmet, G., Ortiz, R., & Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci*, 242, 23-36.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, 67, 48.
- Bawa, K.S. (1980). Evolution of dioecy in flowering plants. *Annu Rev Ecol Evol Syst*, 11, 15-39.
- Beerling, D.J. (1998). *Salix herbacea* L. *J Ecol*, 86, 872-895.
- Begley, D., McCracken, A.R., Dawson, W.M., & Watson, S. (2009). Interaction in short rotation coppice willow, *Salix viminalis* genotype mixtures. *Biomass Bioenerg*, 33, 163-173.
- Beismann, H., Barker, J.H.A., Karp, A., & Speck, T. (1997). AFLP analysis sheds light on distribution of two *Salix* species and their hybrid along a natural gradient. *Mol Ecol*, 6,

989-993.

- Bergero, R., Charlesworth, D., Filatov, D.A., & Moore, R.C. (2008). Defining regions and rearrangements of the *Silene latifolia* Y chromosome. *Genetics*, 178, 2045-2053.
- Berlin, S., Ghelardini, L., Bonosi, L., Weih, M., & Rönnerberg-Wästljung, A. (2014a). QTL mapping of biomass and nitrogen economy traits in willows (*Salix* spp.) grown under contrasting water and nutrient conditions. *Mol Breeding*, 34, 1987-2003.
- Berlin, S., Trybush, S.O., Fogelqvist, J., Gyllenstrand, N., Hallingbäck, H.R., Åhman, I., et al. (2014b). Genetic diversity, population structure and phenotypic variation in European *Salix viminalis* L. (Salicaceae). *Tree Genet Genomes*, 10, 1595.
- Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. *Bioinformatics*, 27, 2463-2464.
- Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., & Buckler, E.S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23, 2633-2635.
- Brereton, N.J.B., Pitre, F.E., Hanley, S.J., Ray, M.J., Karp, A., & Murphy, R.J. (2010). QTL mapping of enzymatic saccharification in short rotation coppice willow and its independence from biomass yield. *Bioenerg Res*, 3, 251-261.
- Brown, H.P. (1921). *Trees of New York state : Native and naturalized* (Vol. XXI, no. 5). Syracuse, N.Y.: The University.
- Cameron, K.D., Phillips, I.S., Kopp, R.F., Volk, T.A., Maynard, C.A., Abrahamson, L.P., et al. (2008). Quantitative genetics of traits indicative of biomass production and heterosis in 34 full-sib F₁ *Salix eriocephala* families. *Bioenerg Res*, 1, 80-90.
- Carolyn, S.W. and Rundel, P.W. (1979). Sexual dimorphism and resource allocation in male and female shrubs of *Simmondsia chinensis*. *Oecologia*, 44, 34-39.
- Charlesworth, D. (1999). Theories of the evolution of dioecy. In M. A. Geber, T. E. Dawson, & L. F. Delph (Eds.), *Gender and sexual dimorphism in flowering plants* (pp. 33-60). Berlin, Heidelberg: Springer Berlin Heidelberg.

- Charlesworth, D. (2002). Plant sex determination and sex chromosomes. *Heredity*, 88, 94-101.
- Charlesworth, D. (2013). Plant sex chromosome evolution. *J Exp Bot*, 64, 405-420.
- Charlesworth, D. (2015). Plant contributions to our understanding of sex chromosome evolution. *New Phytol*, 208, 52-65.
- Charlesworth, D. and Guttman, D.S. (1999). The evolution of dioecy and plant sex chromosome systems. In C. Ainsworth (Ed.), *Sex determination in plants* (pp. 25-49). Oxford, UK: BIOS.
- Charlesworth, D. and Willis, J.H. (2009). The genetics of inbreeding depression. *Nat Rev Genet*, 10, 783-796.
- Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G., & Zanne, A.E. (2009). Towards a worldwide wood economics spectrum. *Eco Lett*, 12, 351-366.
- Chaves, A.L., Vergara, C.E., & Mayer, J.E. (1995). Dichloromethane as an economic alternative to chloroform in the extraction of DNA from plant tissues. *Plant Mol Biol Report*, 13, 18-25.
- Che-Castaldo, C., Crisafulli, C.M., Bishop, J.G., & Fagan, W.F. (2015). What causes female bias in the secondary sex ratios of the dioecious woody shrub *Salix sitchensis* colonizing a primary successional landscape? *Am J Bot*, 102, 1309-1322.
- Chen, Y., Wang, T., Fang, L., Li, X., & Yin, T. (2016). Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. *PLoS ONE*, 11, e0147671.
- Collard, B.C.Y. and Mackill, D.J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philos T R Soc B*, 363, 557-572.
- Collinson, M.E. (1992). The early fossil history of Salicaceae: A brief review. *P Roy Soc Edinb B*, 98B, 155-167.
- Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D., et al. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature*, 508, 488-493.

- Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J.M., & Chen, C. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: (Bethesda)*, 3, 1903-1926.
- Crossa, J. and Federer, W.T. (2012). I.4 Screening experimental designs for quantitative trait loci, association mapping, genotype-by environment interaction, and other investigations. *Front Physiol*, 3, 156.
- Darwin, C. (1877). *The different forms of flowers on plants of the same species*. London, UK: Cambridge University Press.
- Davies, G. and Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure*, 3, 853-859.
- Dawson, T.E. and Geber, M.A. (1999). Sexual dimorphism in physiology and morphology. In M. A. Geber, T. E. Dawson, & L. F. Delph (Eds.), *Gender and sexual dimorphism in flowering plants* (pp. 175-215). Berlin, Heidelberg: Springer Berlin Heidelberg.
- DeFaveri, J., Viitaniemi, H., Leder, E., & Merila, J. (2013). Characterizing genic and nongenic molecular markers: Comparison of microsatellites and SNPs. *Mol Ecol Resour*, 13, 377-392.
- Dekkers, J.C. (2007). Marker-assisted selection for commercial crossbred performance. *J Anim Sci*, 85, 2104-2114.
- Delph, L.F. (1999). Sexual dimorphism in life history. In M. A. Geber, T. E. Dawson, & L. F. Delph (Eds.), *Gender and sexual dimorphism in flowering plants* (pp. 149-173). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Delph, L.F. (2009). Sex Allocation: Evolution to and from dioecy. *Curr Biol*, 19, R249-R251.
- Denis, M. and Bouvet, J.-M. (2011). Genomic selection in tree breeding: testing accuracy of prediction models including dominance effect. *BMC Proc*, 5, 1.
- Denis, M., Favreau, B., Ueno, S., Camus-Kulandaivelu, L., Chaix, G., Gion, J.M., et al. (2013). Genetic variation of wood chemical traits and association with underlying genes in *Eucalyptus urophylla*. *Tree Genet Genomes*, 9, 927-942.

- Dickerson, J. (2002). Purple osier willow *Salix purpurea* L.: Plant fact sheet. Syracuse, NY, USA: USDA Natural Resources Conservation Service.
- Dickmann, D. and Kuzovkina, J. (2008). Poplars and willows in the world *Poplars and willows in the world, Meeting the needs of society and the environment. International Poplar Commission* (pp. 8-91). Rome, Italy: FAO.
- Douhovnikoff, V. and Dodd, R.S. (2003). Intra-clonal variation and a similarity threshold for identification of clones: Application to *Salix exigua* using AFLP molecular markers. *Theor Appl Genet*, 106, 1307-1315.
- Douhovnikoff, V., McBride, J.R., & Dodd, R.S. (2005). *Salix exigua* clonal growth and population dynamics in relation to disturbance regime variation. *Ecology*, 86, 446-452.
- Du, Q., Wang, L., Yang, X., Gong, C., & Zhang, D. (2015). *Populus* endo-beta-1,4-glucanases gene family: genomic organization, phylogenetic analysis, expression profiles and association mapping. *Planta*, 241, 1417-1434.
- Dudley, L.S. (2006). Ecological correlates of secondary sexual dimorphism in *Salix glauca* (Salicaceae). *Am J Bot*, 93, 1775-1783.
- Dudley, L.S. and Galen, C. (2007). Stage-dependent patterns of drought tolerance and gas exchange vary between sexes in the alpine willow, *Salix glauca*. *Oecologia*, 153, 1-9.
- Earl, D. and vonHoldt, B. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour*, 4, 359-361.
- Edwards, A.W.F. (2000). Carl Düsing (1884) on the regulation of the sex-ratio. *Theor Popul Biol*, 58, 255-257.
- Ehlers, B.K. and Bataillon, T. (2007). 'Inconstant males' and the maintenance of labile sex expression in subdioecious plants. *New Phytol*, 174, 194-211.
- Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6, e19379.

- Emanuelli, F., Lorenzi, S., Grzeskowiak, L., Catalano, V., Stefanini, M., Troggio, M., et al. (2013). Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol*, *13*, 39.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol*, *14*, 2611–2620.
- Evans, L.M., Slavov, G.T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., et al. (2014). Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet*, *46*, 1089-1096.
- Fabio, E.S., Volk, T.A., Miller, R.O., Serapiglia, M.J., Gauch, H.G., Van Rees, K.C., et al. (2016). Genotype by environment interactions analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy*. Retrieved from doi:10.1111/gcbb.12344
- Fahrenkrog, A.M., Neves, L.G., Resende, M.F.R., Vazquez, A.I., de los Campos, G., Dervinis, C., et al. (2016). Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytol*. Retrieved from doi:10.1111/nph.14154
- Falush, D., Stephens, M., & Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567-1587.
- Farmer, R.E. (1964). Sex ratio and sex-related characteristics in eastern cottonwood. *Silvae Genet*, *13*, 116-118.
- Fechter, I., Hausmann, L., Daum, M., Sorensen, T.R., Viehover, P., Weisshaar, B., et al. (2012). Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol Genet Genomics*, *287*, 247-259.
- Field, D.L., Pickup, M., & Barrett, S.C.H. (2013a). Comparative analyses of sex-ratio variation in dioecious flowering plants. *Evolution*, *67*, 661-672.
- Field, D.L., Pickup, M., & Barrett, S.C.H. (2013b). Ecological context and metapopulation dynamics affect sex-ratio variation among dioecious plant populations. *Ann Bot*, *111*, 917-923.
- Filatov, D.A. (2015). Homomorphic plant sex chromosomes are coming of age. *Mol Ecol*, *24*,

3217-3219.

Fisher, R.A. (1930). *The genetical theory of natural selection: a complete variorum edition*. London, UK: Oxford University Press.

Frey, B.J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*, 972-976.

Garnett, T., Appleby, M.C., Balmford, A., Bateman, I.J., Benton, T.G., Bloomer, P., et al. (2013). Sustainable intensification in agriculture: Premises and policies. *Science*, *341*, 33-34.

Gaut, B.S. and Long, A.D. (2003). The lowdown on linkage disequilibrium. *Plant Cell*, *15*, 1502-1506.

Geraldes, A., DiFazio, S.P., Slavov, G.T., Ranjan, P., Muchero, W., Hannemann, J., et al. (2013). A 34K SNP genotyping array for *Populus trichocarpa*: Design, application to the study of natural populations and transferability to other *Populus* species. *Mol Ecol Resour*, *13*, 306-323.

Geraldes, A., Hefer, C.A., Capron, A., Kolosova, N., Martinez-Nuñez, F., Soolanayakanahally, R.Y., et al. (2015). Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Mol Ecol*, *24*, 3243-3256.

Ghelardini, L., Berlin, S., Weih, M., Lagercrantz, U., Gyllenstrand, N., & Rönnerberg-Wästljung, A.-C. (2014). Genetic architecture of spring and autumn phenology in *Salix*. *BMC Plant Biol*, *14*, 31.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., et al. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE*, *9*, e90346.

González, E., González-Sanchis, M., Cabezas, Á., Comín, F.A., & Muller, E. (2010). Recent changes in the riparian forest of a large regulated mediterranean river: Implications for management. *Environ Manage*, *45*, 669-681.

Goodwin, S., McPherson, J.D., & McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*, 333-351.

- Gramlich, S., Sagmeister, P., Dullinger, S., Hadacek, F., & Horandl, E. (2016). Evolution *in situ*: Hybrid origin and establishment of willows (*Salix* L.) on alpine glacier forefields. *Heredity*, 116, 531-541.
- Grattapaglia, D. and Resende, M. (2011). Genomic selection in forest tree breeding. *Tree Genet Genomes*, 7, 241-255.
- Gupta, H., Agrawal, P., Mahajan, V., Bisht, G., Kumar, A., Verma, P., et al. (2009). Quality protein maize for nutritional security: Rapid development of short duration hybrids through molecular marker assisted breeding. *Curr Sci*, 96, 230-237.
- Hallingbäck, H.R., Fogelqvist, J., Powers, S.J., Turrión-Gómez, J., Rossiter, R., Amey, J., et al. (2015). Association mapping in *Salix viminalis* L. (Salicaceae) - identification of candidate genes associated with growth and phenology. *GCB Bioenergy*, 8, 670-685.
- Hanley, S.J., Barker, J.B., Van Ooijen, J.V.O., Aldam, C.A., Harris, S.H., Åhman, I.Å., et al. (2002). A genetic linkage map of willow (*Salix viminalis*) based on AFLP and microsatellite markers. *Theor Appl Genet*, 105, 1087-1096.
- Hanley, S.J. and Karp, A. (2013). Genetic strategies for dissecting complex traits in biomass willows (*Salix* spp.). *Tree Physiol*, 34, 1167-1180.
- Hanley, S.J., Pei, M.H., Powers, S.J., Ruiz, C., Mallott, M.D., Barker, J.H.A., et al. (2011). Genetic mapping of rust resistance loci in biomass willow. *Tree Genet Genomes*, 7, 597-608.
- Hardig, T.M., Brunsfeld, S.J., Fritz, R.S., Morgan, M., & Orians, C.M. (2000). Morphological and molecular evidence for hybridization and introgression in a willow (*Salix*) hybrid zone. *Mol Ecol*, 9, 9-24.
- He, S., Schulthess, A.W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., et al. (2016). Genomic selection in a commercial winter wheat population. *Theor Appl Genet*, 129, 641-651.
- Heffner, E.L., Lorenz, A.J., Jannink, J.-L., & Sorrells, M.E. (2010). Plant breeding with genomic selection: Gain per unit time and cost. *Crop Sci*, 50, 1681-1690.
- Heffner, E.L., Sorrells, M.E., & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci*, 49, 1-12.

- Heller, M.C., Keoleian, G.A., & Volk, T.A. (2003). Life cycle assessment of a willow bioenergy cropping system. *Biomass Bioenerg*, 25, 147-165.
- Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6, 95 - 108.
- Hou, J., Ye, N., Zhang, D., Chen, Y., Fang, L., Dai, X., et al. (2015). Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. *Sci Rep*, 5, 1-6.
- Houle, D., Govindaraju, D.R., & Omholt, S. (2010). Phenomics: The next challenge. *Nat Rev Genet*, 11, 855-866.
- Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet*, 42, 961-967.
- Hughes, F.M.R., Johansson, M., Xiong, S., Carlborg, E., Hawkins, D., Svedmark, M., et al. (2009). The influence of hydrological regimes on sex ratios and spatial segregation of the sexes in two dioecious riparian shrub species in northern Sweden. *Plant Ecol*, 208, 77-92.
- Hultine, K.R., Grady, K.C., Wood, T.E., Shuster, S.M., Stella, J.C., & Whitham, T.G. (2016). Climate change perils for dioecious plant species. *Nat Plants*, 2, 16109.
- Iida, Y., Poorter, L., Sterck, F.J., Kassim, A.R., Kubo, T., Potts, M.D., et al. (2012). Wood density explains architectural differentiation across 145 co-occurring tropical tree species. *Funct Ecol*, 26, 274-282.
- Ingvarsson, P.K. (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae). *Genetics*, 169, 945-953.
- Ingvarsson, P.K., Garcia, M.V., Luquez, V., Hall, D., & Jansson, S. (2008). Nucleotide polymorphism and phenotypic associations within and around the *phytochrome B2* Locus in European aspen (*Populus tremula*, Salicaceae). *Genetics*, 178, 2217-2226.
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, 436, 793-800.
- Ioannidis, J.P., Thomas, G., & Daly, M.J. (2009). Validating, augmenting and refining genome-

- wide association signals. *Nat Rev Genet*, 10, 318-329.
- Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449, 463-467.
- Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, 15, 740.
- Jena, K.K. and Mackill, D.J. (2008). Molecular markers and their use in marker-assisted selection in rice *Crop Sci*, 48, 1266-1276.
- Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27, 3070-3071.
- Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet*, 11, 94.
- Jones, P. (1997). Microclimate in open-top chambers: Implications for predicting climate change effects on rice production. *Transactions of the ASAE*, 40, 739.
- Julkunen-Tiitto, R. (1996). Defensive efforts of *Salix myrsinifolia* plantlets in photomixotrophic culture conditions: The effect of sucrose, nitrogen and pH on the phytomass and secondary phenolic accumulation. *Écoscience*, 3, 297-303.
- Kamvar, Z.N., Tabima, J.F., & Grünwald, N.J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281.
- Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42, 348-354.
- Karp, A., Hanley, S., Trybush, S., Macalpine, W., Pei, M.H., & Shield, I. (2011). Genetic improvement of willow for bioenergy and biofuels. *J Integr Plant Biol*, 53, 151-165.
- Karrenberg, S., Kollmann, J., & Edwards, P.J. (2002). Pollen vectors and inflorescence morphology in four species of *Salix*. *Pl Syst Evol*, 235, 181-188.

- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647-1649.
- Kenaley, S.C., Smart, L.B., & Hudler, G.W. (2014). Genetic evidence for three discrete taxa of *Melampsora* (*Pucciniales*) affecting willows (*Salix* spp.) in New York State. *Fungal Biol*, 118, 704-720.
- Kersten, B., Pakull, B., Groppe, K., Lueneburg, J., & Fladung, M. (2014). The sex-linked region in *Populus tremuloides* Turesson 141 corresponds to a pericentromeric region of about two million base pairs on *P. trichocarpa* chromosome 19. *Plant Biol*, 16, 411-418.
- Khan, M.A. and Korban, S.S. (2012). Association mapping in forest trees and fruit crops. *J Exp Bot*, 63, 4045-4060.
- Kiddle, S.J., Windram, O.P.F., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., et al. (2010). Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*. *Bioinformatics*, 26, 355-362.
- Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.S., & Paterson, A.H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci*, 242, 14-22.
- Kim, S.H., Gao, F., Bhattacharjee, S., Adiasor, J.A., Nam, J.C., & Gassmann, W. (2010). The *Arabidopsis* resistance-like gene SNC1 is activated by mutations in SRFR1 and contributes to resistance to the bacterial effector AvrRps4. *PLoS Pathog*, 6, e1001172.
- Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., & Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inferences across *K*. *Mol Ecol Resour*, 15, 1179-1191.
- Kruijer, W., Boer, M.P., Malosetti, M., Flood, P.J., Engel, B., Kooke, R., et al. (2015). Marker-based estimation of heritability in immortal populations. *Genetics*, 199, 379-398.
- Kuzovkina, Y. and Quigley, M.F. (2005). Willows beyond wetlands: Uses of *Salix* L. species for environmental projects. *Water Air Soil Pollut*, 162, 183-204.
- Kuzovkina, Y.A., Weih, M., Romero, M.A., Charles, J., Hust, S., McIvor, I., et al. (2008). *Salix*: Botany and Global Horticulture *Horticultural Reviews* (pp. 447-489). Hoboken, NJ: John

Wiley & Sons, Inc.

- Kwon, S.I., Kim, S.H., Bhattacharjee, S., Noh, J.J., & Gassmann, W. (2009). SRFR1, a suppressor of effector-triggered immunity, encodes a conserved tetratricopeptide repeat protein with similarity to transcriptional repressors. *Plant J*, 57, 109-119.
- Lascoux, M., Thorsén, J., & Gullberg, U. (1996). Population structure of a riparian willow species, *Salix viminalis* L. *Genet Res*, 68, 45-54.
- Lauron-Moreau, A., Pitre, F.E., Argus, G.W., Labrecque, M., & Brouillet, L. (2015). Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS ONE*, 10, e0121965.
- Lawrence, G.J., Anderson, P.A., Dodds, P.N., & Ellis, J.G. (2010). Relationships between rust resistance genes at the *M* locus in flax. *Mol Plant Pathol*, 11, 19-32.
- Leone, M., Sumedha, & Weigt, M. (2007). Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics*, 23, 2708-2715.
- Lewis, D. (1942). The evolution of sex in flowering plants. *Biol Rev*, 17, 46-67.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-1760.
- Lin, J., Gibbs, J.P., & Smart, L.B. (2009). Population genetic structure of native versus naturalized sympatric shrub willows (*Salix*: Salicaceae). *Am J Bot*, 96, 771-785.
- Lin, J. and Zsuffa, L. (1993). Quantitative genetic parameters for seven characteristics in a clonal test of *Salix eriocephala*. I. Clonal variation, clone environment interactions, heritability, and genetic gains. *Silvae Genet*, 42, 41-46.
- Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28, 2397-2399.
- Liu, N., Nielsen, H.K., Jørgensen, U., & Lærke, P.E. (2015). Sampling procedure in a willow plantation for chemical elements important for biomass combustion quality. *Fuel*, 142, 283-288.

- Lloyd, D.G. (1974). Female-predominant sex ratios in angiosperms. *Heredity*, 32, 35-44.
- Lloyd, D.G. (1982). Selection of combined versus separate sexes in seed plants. *Am Nat*, 120, 571-585.
- Lloyd, D.G. and Webb, C.J. (1977). Secondary sex characters in plants. *Bot Rev*, 43, 177-216.
- Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., & Casler, M.D. (2013). Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genet*, 9, e1003215.
- Mackay, T.F.C., Stone, E.A., & Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*, 10, 565-577.
- McCracken, A.R. and Dawson, W.M. (2003). Rust disease (*Melampsora epitea*) of willow (*Salix* spp.) grown as short rotation coppice (SRC) in inter- and intra-species mixtures. *Ann Appl Biol*, 143, 381-393.
- McKown, A.D., Klápště, J., Guy, R.D., Geraldès, A., Porth, I., Hannemann, J., et al. (2014). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol*, 203, 535-553.
- Meikle, R.D. (1992). British willows; Some hybrids and some problems. *P Roy Soc Edinb B*, 98, 13-20.
- Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., & Koornneef, M. (1998). *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, 282, 662-682.
- Meuwissen, T.H., Hayes, B.J., & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819-1829.
- Miller, R.O. and Bender, B.A. (2010). *Four years of herbicide trials for shrub willow biomass production systems in the Upper Peninsula of Michigan*. Paper presented at the Short Rotation Woody Crops International Conference, Syracuse University, Syracuse, NY. Oral Presentation retrieved from http://www.esf.edu/outreach/pd/2010/srwc/documents/rmiller_fouryearsofherbicide.pdf
- Ming, R., Bendahmane, A., & Renner, S.S. (2011). Sex chromosomes in land plants. *Annu Rev*

- Plant Biol*, 62, 485-514.
- Ming, R., Wang, J., Moore, P.H., & Paterson, A.H. (2007). Sex chromosomes in flowering plants. *Am J Bot*, 94, 141-150.
- Moggridge, H.L. and Gurnell, A.M. (2009). Controls on the sexual and asexual regeneration of Salicaceae along a highly dynamic, braided river system. *Aquat Sci*, 71, 305-317.
- Montes, J.M., Melchinger, A.E., & Reif, J.C. (2007). Novel throughput phenotyping platforms in plant genetic studies. *Trends Plant Sci*, 12, 433-436.
- Montgomery, S.B., Goode, D., Kvikstad, E., Albers, C.A., Zhang, Z., Mu, X.J., et al. (2013). The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*.
- Moritz, K.K., Björkman, C., Parachnowitsch, A.L., & Stenberg, J.A. (2016). Female *Salix viminalis* are more severely infected by *Melampsora* spp. but neither sex experiences associational effects. *Ecol Evol*, 6, 1154-1162.
- Morris, G.P., Grabowski, P.P., & Borevitz, J.O. (2011). Genomic diversity in switchgrass (*Panicum virgatum*): From the continental scale to a dune landscape. *Mol Ecol*, 20, 4938-4952.
- Mosseler, A., Major, J.E., & Labrecque, M. (2014). Genetic by environment interactions of two North American *Salix* species assessed for coppice yield and components of growth on three sites of varying quality. *Trees*, 28, 1-11.
- Mosseler, A. and Zsuffa, L. (1989). Sex expression and sex ratios in intra-and interspecific hybrid families of *Salix* L. *Silvae Genet*, 38, 12-17.
- Myers-Smith, I.H. and Hik, D.S. (2012). Uniform female-biased sex ratios in alpine willows. *Am J Bot*, 99, 1243-1248.
- Nagamitsu, T., Hoshikawa, T., Kawahara, T., Barkalov, V.Y., & Sabirov, R.N. (2014). Phylogeography and genetic structure of disjunct *Salix arbutifolia* populations in Japan. *Popul Ecol*, 56, 539-549.
- Obeso, J. (2002). The costs of reproduction in plants. *New Phytol*, 155, 321-348.

- Olson, M.S., Levsen, N., Soolanayakanahally, R.Y., Guy, R.D., Schroeder, W.R., Keller, S.R., et al. (2013). The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Mol Ecol*, 22, 1214-1230.
- Pakull, B., Groppe, K., Mecucci, F., Gaudet, M., Sabatti, M., & Fladung, M. (2011). Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a *Populus tremula* × *Populus tremuloides* cross. *Can J For Res*, 41, 245-253.
- Pakull, B., Groppe, K., Meyer, M., Markussen, T., & Fladung, M. (2009). Genetic linkage mapping in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). *Tree Genet Genomes*, 5, 505-515.
- Paradis, E. (2010). pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26, 419-420.
- Paterson, A., Bowers, J., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551 - 556.
- Pauley, S.S. (1948). Sex and vigor in *Populus*. *Science*, 108, 302-303.
- Pei, M.H., Lindegaard, K., Ruiz, C., & Bayon, C. (2008). Rust resistance of some varieties and recently bred genotypes of biomass willows. *Biomass Bioenerg*, 32, 453-459.
- Perdereau, A., Kelleher, C., Douglas, G., & Hodgkinson, T. (2014). High levels of gene flow and genetic diversity in Irish populations of *Salix caprea* L. inferred from chloroplast and nuclear SSR markers. *BMC Plant Biol*, 14, 1-12.
- Peterson, G., Dong, Y., Horbach, C., & Fu, Y.-B. (2014). Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, 6, 665-680.
- Peto, F.H. (1938). Cytology of poplar species and natural hybrids. *Can J Res*, 16c, 445-455.
- Petzold, A., Pfeiffer, T., Jansen, F., Eusemann, P., & Schnittler, M. (2012). Sex ratios and clonal growth in dioecious *Populus euphratica* Oliv., Xinjiang Prov., Western China. *Trees*, 27, 729-744.
- Phillips, I.S. (2002). *Quantitative genetics of traits predictive of biomass yield in first- and second-generation Salix eriocephala*. (M.S.), SUNY College of Environmental Science

and Forestry, Syracuse, NY.

Poland, J.A. and Rife, T.W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome*, 5, 92-102.

Porth, I., Klapšte, J., Skyba, O., Hannemann, J., McKown, A.D., Guy, R.D., et al. (2013). Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytol*, 200, 710-726.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38, 904-909.

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., & Donnelly, P. (2000b). Association mapping in structured populations. *Am J Hum Genet*, 67, 170-181.

Pucholt, P., Rönnerberg-Wästljung, A.-C., & Berlin, S. (2015). Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity*, 114, 575-583.

Rasband, W.S. (1997-2016). ImageJ. from U. S. National Institutes of Health
<http://imagej.nih.gov/ij/>

Rechinger, K.H. (1992). *Salix* taxonomy in Europe – Problems, interpretations, observations. *P Roy Soc Edinb B*, 98, 1-12.

Reisch, C., Schurm, S., & Poschlod, P. (2007). Spatial genetic structure and clonal diversity in an alpine population of *Salix herbacea* (Salicaceae). *Ann Bot*, 99, 647-651.

Renner, S.S. (2014). The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. *Am J Bot*, 101, 1588-1596.

Resende, M.D.V., Resende, M.F.R., Sansaloni, C.P., Petroli, C.D., Missiaggia, A.A., Aguiar, A.M., et al. (2012). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest

- trees. *New Phytol*, 194, 116-128.
- Robinson, K.M., Delhomme, N., Mähler, N., Schiffthaler, B., Önskog, J., Albrechtsen, B.R., et al. (2014). *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. *BMC Plant Biol*, 14, 1-14.
- Rönnberg-Wästljung, A.-C. (2001). Genetic structure of growth and phenological traits in *Salix viminalis*. *Can J For Res*, 31, 276-282.
- Rönnberg-Wästljung, A.-C., Samils, B., Tsarouhas, V., & Gullberg, U. (2008). Resistance to *Melampsora larici-epitea* leaf rust in *Salix*: Analyses of quantitative trait loci. *J Appl Genet*, 49, 321-331.
- Rönnberg-Wästljung, A.-C., Tsarouhas, V., Semerikov, V., & Lagercrantz, U. (2003). A genetic linkage map of a tetraploid *Salix viminalis* x *S. dasyclados* hybrid based on AFLP markers. *For Genet*, 10, 185-194.
- Rönnberg-Wästljung, A.C. and Gullberg, U. (1999). Genetics of breeding characters with possible effects on biomass production in *Salix viminalis* (L.). *Theor Appl Genet*, 98, 531-540.
- Rönnberg-Wästljung, A.C., Gullberg, U., & Nilsson, C. (1994). Genetic parameters of growth characters in *Salix viminalis* grown in Sweden. *Can J For Res*, 24, 1960-1969.
- Rosenberg, N.A. (2004). DISTRUCT: A program for the graphical display of population structure. *Mol Ecol Notes*, 4, 137-138.
- Rottenberg, A. (1998). Sex ratio and gender stability in the dioecious plants of Israel. *Bot J Linn Soc*, 128, 137-148.
- Sakai, A., Sasa, A., & Sakai, S. (2006). Do sexual dimorphisms in reproductive allocation and new shoot biomass increase with an increase of altitude? A case of the shrub willow *Salix reinii* (Salicaceae). *Am J Bot*, 93, 988-992.
- Sakai, A.K. and Weller, S.G. (1999). Gender and sexual dimorphism in flowering plants: A review of terminology, biogeographic patterns, ecological correlates, and phylogenetic approaches. In M. A. Geber, T. E. Dawson, & L. F. Delph (Eds.), *Gender and sexual dimorphism in flowering plants* (pp. 1-31). Berlin, Heidelberg: Springer Berlin Heidelberg.

- Salix purpurea* v1.0, DOE-JGI. (2015). Available from DOE-JGI
http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Spurpurea
- Samils, B., Rönnerberg-Wästljung, A.-C., & Stenlid, J. (2011). QTL mapping of resistance to leaf rust in *Salix*. *Tree Genet Genomes*, 7, 1219-1235.
- Sanderson, M.A., Reed, R.L., McLaughlin, S.B., Wulschleger, S.D., Conger, B.V., Parrish, D.J., et al. (1996). Switchgrass as a sustainable bioenergy crop. *Bioresour Technol*, 56, 83-93.
- Sannigrahi, P., Ragauskas, A.J., & Tuskan, G.A. (2010). Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4, 209-226.
- Saska, M.M. and Kuzovkina, Y.A. (2010). Phenological stages of willow (*Salix*). *Ann Appl Biol*, 156, 431-437.
- Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, 8, 552-560.
- Schapiro, A.L., Valpuesta, V., & Botella, M.A. (2006). TPR proteins in plant hormone signaling. *Plant Signaling & Behavior*, 1, 229-230.
- Schneider, C.A., Rasband, W.S., & Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Meth*, 9, 671-675.
- Seeger, J. and Eckhart, V.M. (1996). Evolution of sexual systems and sex allocation in plants when growth and reproduction overlap. *P Roy Soc B-Biol Sci*, 263, 833-841.
- Semerikov, V., Lagercrantz, U., Tsarouhas, V., Rönnerberg-Wästljung, A., Alström-Rapaport, C., & Lascoux, M. (2003). Genetic mapping of sex-linked markers in *Salix viminalis* L. *Heredity*, 91, 293-299.
- Serapiglia, M.J., Cameron, K.D., Stipanovic, A.J., Abrahamson, L.P., Volk, T.A., & Smart, L.B. (2013). Yield and woody biomass traits of novel shrub willow hybrids at two contrasting sites. *Bioenerg Res*, 6, 533-546.
- Serapiglia, M.J., Cameron, K.D., Stipanovic, A.J., & Smart, L.B. (2009). Analysis of biomass composition using high-resolution thermogravimetric analysis and percent bark content

- for the selection of shrub willow bioenergy crop varieties. *Bioenerg Res*, 2, 1-9.
- Serapiglia, M.J., Gouker, F.E., Hart, J.F., Unda, F., Mansfield, S.D., Stipanovic, A.J., et al. (2014a). Ploidy level affects important biomass traits of novel shrub willow (*Salix*) hybrids. *Bioenerg Res*, 8, 259-269.
- Serapiglia, M.J., Gouker, F.E., & Smart, L.B. (2014b). Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biol*, 14, 74.
- Setsuko, S., Nagamitsu, T., & Tomaru, N. (2013). Pollen flow and effects of population structure on selfing rates and female and male reproductive success in fragmented *Magnolia stellata* populations. *BMC Ecology*, 13, 10.
- Shafroth, P.B., Scott, M.L., Friedman, J.M., & Laven, R.D. (1994). Establishment, sex structure and breeding system of an exotic riparian willow, *Salix x rubens*. *Am Midl Nat*, 132, 159-172.
- Shield, I., Macalpine, W., Hanley, S., & Karp, A. (2015). Breeding willow for short rotation coppice energy cropping. In V. M. V. Cruz & D. A. Dierig (Eds.), *Industrial Crops* (Vol. 9, pp. 67-80): Springer New York.
- Singh, A., Ganapathysubramanian, B., Singh, A.K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci*, 21, 110-124.
- Skvortsov, A.K. (1999). *Willows of Russia and adjacent countries. Taxonomical and geographical revision* (I. N. Kadis, Trans. G. W. Argus Ed.). Joensuu, Finland: University of Joensuu.
- Slavov, G.T., DiFazio, S.P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol*, 196, 713-725.
- Slavov, G.T., Leonardi, S., Burczyk, J., Adams, W.T., Strauss, S.H., & Difazio, S.P. (2009). Extensive pollen flow in two ecologically contrasting populations of *Populus trichocarpa*. *Mol Ecol*, 18, 357-373.
- Smart, L., Cameron, K.D., Volk, T.A., & Abrahamson, L.P. (2007). *Breeding, selection and testing of shrub willow as a dedicated energy crop*. Paper presented at the National

Agricultural Biotechnology Council, Ithaca, NY.

<http://www.cabdirect.org/abstracts/20083097496.html?freeview=true#>

- Smart, L., Volk, T.A., Lin, J., Kopp, R.F., Phillips, I.S., Cameron, K.D., et al. (2005). Genetic improvement of shrub willow (*Salix* spp.) crops for bioenergy and environmental applications in the United States. *Unasylva*, 56, 51-55.
- Smart, L.B. and Cameron, K.D. (2008). Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In W. Vermerris (Ed.), *Genetic Improvement of Bioenergy Crops* (pp. 377-396): Springer New York.
- Smart, L.B. and Cameron, K.D. (2012). Shrub Willow. In C. Kole, C. P. Joshi, & D. R. Shonnard (Eds.), *Handbook of Bioenergy Crop Plants* (pp. 687-708). Boca Raton, FL: CRC Press.
- Solberg, T., Sonesson, A., & Woolliams, J. (2008). Genomic selection using different marker types and densities. *J Anim Sci*, 86, 2447-2454.
- Soto-Cerda, B.J., Duguid, S., Booker, H., Rowland, G., Diederichsen, A., & Cloutier, S. (2013). Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. *J Integr Plant Biol*, 56, 75-87.
- Spigler, R.B., Lewers, K.S., Main, D.S., & Ashman, T.L. (2008). Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity*, 101, 507-517.
- Stehlik, I. and Barrett, S.C.H. (2006). Pollination intensity influences sex ratios in dioecious *Rumex nivalis*, a wind-pollinated plant. *Evolution*, 60, 1207-1214.
- Stokes, K.E. (2008). Exotic invasive black willow (*Salix nigra*) in Australia: Influence of hydrological regimes on population dynamics. *Plant Ecol*, 197, 91-105.
- Stolarski, M.J., Szczukowski, S., Tworkowski, J., & Klasa, A. (2011). Willow biomass production under conditions of low-input agriculture on marginal soils. *For Ecol Manage*, 262, 1558-1566.
- Sulima, P., Przyborowski, J.A., & Załuski, D. (2009). RAPD markers reveal genetic diversity in *Salix purpurea* L. *Crop Sci*, 49, 857-863.

- Sulima, P. and Przyborowski Jerzy, A. (2013). Genetic diversity of *Salix purpurea* L. genotypes and interspecific hybrids. *Acta Biol Cracov Bot*, 55, 29-36.
- Sun, G., Ji, Q., Dilcher, D.L., Zheng, S., Nixon, K.C., & Wang, X. (2002). Archaeofractaceae, a new basal angiosperm family. *Science*, 296, 899-904.
- Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT Version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome*, 9, 1-9.
- Tanksley, S.D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, 335, 721-726.
- Taylor, D.R. (1999). Genetics of sex ratio variation among natural populations of a dioecious plant. *Evolution*, 53, 55-62.
- Tester, M. and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science*, 327, 818-822.
- Tharakan, P.J., Volk, T.A., Nowak, C.A., & Ofezu, G.J. (2008). Assessment of canopy structure, light interception, and light-use efficiency of first year regrowth of shrub willow (*Salix* sp.). *Bioenerg Res*, 1, 229-238.
- Thomas, A., Abel, H.J., Di, Y., Faye, L.L., Jin, J., Liu, J., et al. (2011). Effect of linkage disequilibrium on the identification of functional variants. *Genet Epidemiol*, 35, S115-S119.
- Trybush, S.O., Jahodová, Š., Čížková, L., Karp, A., & Hanley, S.J. (2012). High levels of genetic diversity in *Salix viminalis* of the Czech Republic as revealed by microsatellite markers. *Bioenerg Res*, 5, 969-977.
- Tsarouhas, V. (2002). *Genome mapping of quantitative trait loci in Salix with an emphasis on freezing resistance*. (Ph.D. Dissertation), Swedish University of Agricultural Sciences, Uppsala, Sweden.
- Tsarouhas, V., Gullberg, U., & Lagercrantz, U. (2002). An AFLP and RFLP linkage map and quantitative trait locus (QTL) analysis of growth traits in *Salix*. *Theor Appl Genet*, 105, 277-288.

- Tsarouhas, V., Gullberg, U., & Lagercrantz, U.L.F. (2003). Mapping of quantitative trait loci controlling timing of bud flush in *Salix*. *Hereditas*, 138, 172-178.
- TST om-06. (2006). Basic density and moisture content of pulpwood *TAPPI methods 2006*. Technology Park, Atlanta: TAPPI Press.
- Turcotte, J. and Houle, G. (2001). Reproductive costs in *Salix plantifolia* ssp. *plantifolia* in subarctic Québec, Canada. *Écoscience*, 8, 506-512.
- Tuskan, G.A., DiFazio, S., Faivre-Rampant, P., Gaudet, M., Harfouche, A., Jorge, V., et al. (2012). The obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: A retrospective working hypothesis. *Tree Genet Genomes*, 8, 559-571.
- Tuskan, G.A. and DiFazio, S. and Jansson, S. and Bohlmann, J. and Grigoriev, I. and Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313, 1596-1604.
- Tuskan, G.A., DiFazio, S.P., & Teichmann, T. (2004). Poplar genomics is getting popular: The impact of the poplar genome project on tree research. *Plant Biol*, 6, 2-4.
- U.S. Department of Agriculture-Farm Service Agency. (2012). USDA Announces Additional 9,000 Acres for Non-Food Energy Crop Production [Press release]. Retrieved from http://www.fsa.usda.gov/FSA/printapp?fileName=nr_20120613_rel_0195.html&newsType=newsrel
- U.S. Department of Energy. (2016). 2016 Billion-ton study report: Advancing domestic resources for a thriving bioeconomy. In M. H. Langholtz, B. Stokes, & Eaton, LM (Eds.), (Vol. 1: Economic availability of feedstocks. Langholtz, M.H., Stokes, B.J., Eaton, L.M., ORNL/TM-2016/160). Oak Ridge, TN: Oak Ridge National Laboratory.
- Ueno, N., Suyama, Y., & Seiwa, K. (2007). What makes the sex ratio female-biased in the dioecious tree *Salix sachalinensis*? *J Ecol*, 95, 951-959.
- Urbanowicz, B.R., Bennett, A.B., Del Campillo, E., Catala, C., Hayashi, T., Henrissat, B., et al. (2007). Structural organization and a standardized nomenclature for plant endo-1,4-beta-glucanases (cellulases) of glycosyl hydrolase family 9. *Plant Physiol*, 144, 1693-1696.
- Urrestarazu, J., Miranda, C., Santesteban, L., & Royo, J. (2012). Genetic diversity and structure of local apple cultivars from Northeastern Spain assessed by microsatellite markers. *Tree*

Genet Genomes, 8, 1163-1180.

- Valentine, J., Clifton-Brown, J., Hastings, A., Robson, P., Allison, G., & Smith, P. (2012). Food vs. fuel: The use of land for lignocellulosic 'next generation' energy crops that minimize competition with primary food production. *GCB Bioenergy*, 4, 1-19.
- van Buijtenen, J.P. and Einspahr, D.W. (1959). Note on the presence of sex chromosomes in *Populus tremuloides*. *Botanical Gazette*, 121, 60-61.
- Vega-Frutis, R., Munguía-Rosas, M.A., Varga, S., & Kytöviita, M.-M. (2013). Sex-specific patterns of antagonistic and mutualistic biotic interactions in dioecious and gynodioecious plants. *Perspect Plant Ecol Evol Syst*, 15, 45-55.
- Volk, T.A., Abrahamson, L.P., Cameron, K.D., Castellano, P., Corbin, T., Fabio, E., et al. (2011). Yields of willow biomass crops across a range of sites in North America. *Asp Appl Biol*, 112, 67-74.
- Wang, J., Na, J.K., Yu, Q., Gschwend, A.R., Han, J., Zeng, F., et al. (2012). Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc Natl Acad Sci USA*, 109, 13710-13715.
- Wang, Q., Tian, F., Pan, Y., Buckler, E.S., & Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PLoS ONE*, 9, e107684.
- Weih, M. (2009). Genetic and environmental variation in spring and autumn phenology of biomass willows (*Salix* spp.): Effects on shoot growth and nitrogen economy. *Tree Physiol*, 29, 1479-1490.
- Westergaard, M. (1958). The mechanism of sex determination in dioecious flowering plants. *Adv Genet*, 9, 217-281.
- Wezel, A., Casagrande, M., Celette, F., Vian, J.-F., Ferrer, A., & Peigné, J. (2014). Agroecological practices for sustainable agriculture. A review. *Agron Sustain Dev*, 34, 1-20.
- White, J.W., Andrade-Sanchez, P., Gore, M.A., Bronson, K.F., Coffelt, T.A., Conley, M.M., et al. (2012). Field-based phenomics for plant genetics research. *Field Crops Res*, 133, 101-112.

- Winter, D.J. (2012). mmod: An R library for the calculation of population differentiation statistics. *Mol Ecol Resour*, 12, 1158-1160.
- Wolf, D.E., Satkoski, J.A., White, K., & Rieseberg, L.H. (2001). Sex determination in the androdioecious plant *Datisca glomerata* and its dioecious sister species *D. cannabina*. *Genetics*, 159, 1243-1257.
- Wu, J., Nyman, T., Wang, D.-C., Argus, G.W., Yang, Y.-P., & Chen, J.-H. (2015). Phylogeny of *Salix* subgenus *Salix* s.l. (Salicaceae): delimitation, biogeography, and reticulate evolution. *BMC Evol Biol*, 15, 1-13.
- Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., McMullen, M.D., & Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE*, 4, e8451.
- Yang, H., Tao, Y., Zheng, Z., Li, C., Sweetingham, M.W., & Howieson, J.G. (2012). Application of next-generation sequencing for rapid marker development in molecular plant breeding: A case study on anthracnose disease resistance in *Lupinus angustifolius* L. *BMC Genomics*, 13, 318.
- Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., & Nyholt, D.R. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42, 565-569.
- Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., & Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46, 100-106.
- Yang, L., Gong, F., Xiong, E., & Wang, W. (2015). Proteomics: A promising tool for research on sex-related differences in dioecious plants. *Front Plant Sci*, 6, 954.
- Yang, S., Fresnedo-Ramírez, J., Wang, M., Cote, L., Schweitzer, P., Barba, P., et al. (2016). A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: a case study for marker-assisted selection in grapevine. *Hort Res*, 3, 16002.
- Zhang, Z., Ersoz, E., Lai, C., Todhunter, R., Tiwari, H., Gore, M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet*, 42, 355 - 360.
- Zhou, L. and Holliday, J.A. (2012). Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, 13, 703.

- Zhou, Q., Zhang, J., Bachtrog, D., An, N., Huang, Q., Jarvis, E.D., et al. (2014). Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science*, 346, 1333-1334.
- Zhu, C., Gore, M., Buckler, E.S., & Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome*, 1, 5-20.
- Zsuffa, L. (1990). Genetic-improvement of willows for energy plantations. *Biomass*, 22, 35-47.

APPENDIX TO CHAPTER 2

Table A2.1 Clone ID and source information for 267 *S. purpurea* genotypes

Clone ID	Epithet	Longitude	Latitude	State / Province	Country	Region	Source
94001		-75.6333	43.2167	New York	USA	Northeastern USA	Natural accession
94002		-75.6333	43.2167	New York	USA	Northeastern USA	Natural accession
94003		-75.6333	43.2167	New York	USA	Northeastern USA	Natural accession
94004		-75.6333	43.2167	New York	USA	Northeastern USA	Natural accession
94005		-75.6333	43.2167	New York	USA	Northeastern USA	Natural accession
94006		-75.6333	43.2168	New York	USA	Northeastern USA	Natural accession
94009		-75.8500	42.8667	New York	USA	Northeastern USA	Natural accession
94011		-75.8500	42.8667	New York	USA	Northeastern USA	Natural accession
94012		-75.8500	42.8668	New York	USA	Northeastern USA	Natural accession
94013		-75.8500	42.8668	New York	USA	Northeastern USA	Natural accession
94014		-75.8500	42.8668	New York	USA	Northeastern USA	Natural accession
94015		-75.8500	42.8668	New York	USA	Northeastern USA	Natural accession
95001		-75.8167	42.8500	New York	USA	Northeastern USA	Natural accession
95002		-75.7667	42.8506	New York	USA	Northeastern USA	Natural accession
95005		-75.7917	42.8333	New York	USA	Northeastern USA	Natural accession
95026		-73.5500	41.7717	New York	USA	Northeastern USA	Natural accession
95038		-78.1333	42.8167	New York	USA	Northeastern USA	Natural accession
95042		-78.1333	42.8167	New York	USA	Northeastern USA	Natural accession
95049		-79.2333	42.3667	New York	USA	Northeastern USA	Natural accession
95057		-78.2500	42.3000	New York	USA	Northeastern USA	Natural accession
95058		-78.2500	42.3000	New York	USA	Northeastern USA	Natural accession
95071		-76.8333	42.3333	New York	USA	Northeastern USA	Natural accession
00-01-001		-75.5433	43.3595	New York	USA	Northeastern USA	Natural accession
00-01-003		-75.5449	43.3598	New York	USA	Northeastern USA	Natural accession
00-01-004		-75.5460	43.3599	New York	USA	Northeastern USA	Natural accession
00-01-009		-75.5405	43.3998	New York	USA	Northeastern USA	Natural accession
00-01-011		-75.5405	43.3999	New York	USA	Northeastern USA	Natural accession
00-01-014		-75.6795	43.1677	New York	USA	Northeastern USA	Natural accession

Table A2.1 (Continued)

00-01-034	-76.1906	42.7892	New York	USA	Northeastern USA	Natural accession
00-01-085	-76.0405	42.8951	New York	USA	Northeastern USA	Natural accession
00-01-086	-76.0373	42.8940	New York	USA	Northeastern USA	Natural accession
00-01-088	-76.0048	42.8249	New York	USA	Northeastern USA	Natural accession
00-01-089	-75.9515	42.8388	New York	USA	Northeastern USA	Natural accession
00-01-091	-75.6125	42.7447	New York	USA	Northeastern USA	Natural accession
00-01-094	-75.5840	42.7333	New York	USA	Northeastern USA	Natural accession
00-01-095	-75.4431	42.5531	New York	USA	Northeastern USA	Natural accession
00-01-098	-75.4238	42.4664	New York	USA	Northeastern USA	Natural accession
00-01-101	-75.6709	42.9028	New York	USA	Northeastern USA	Natural accession
00-01-102	-75.6803	42.8674	New York	USA	Northeastern USA	Natural accession
00-01-103	-75.6442	43.8010	New York	USA	Northeastern USA	Natural accession
00-01-104	-75.4412	43.6756	New York	USA	Northeastern USA	Natural accession
00-01-105	-75.4195	43.5701	New York	USA	Northeastern USA	Natural accession
00-01-106	-75.4933	43.3657	New York	USA	Northeastern USA	Natural accession
00-22-002	Location not recorded		North Carolina	USA	Southern USA	Natural accession
01-01-001	-76.0504	43.0436	New York	USA	Northeastern USA	Natural accession
01-01-028	-76.1316	42.7503	New York	USA	Northeastern USA	Natural accession
01-01-029	-75.9095	42.9065	New York	USA	Northeastern USA	Natural accession
01-01-030	-75.9104	42.9061	New York	USA	Northeastern USA	Natural accession
01-01-031	-75.9080	42.9064	New York	USA	Northeastern USA	Natural accession
01-01-032	-75.8321	42.8903	New York	USA	Northeastern USA	Natural accession
01-01-034	-75.8253	42.8864	New York	USA	Northeastern USA	Natural accession
01-01-036	-75.8185	42.8539	New York	USA	Northeastern USA	Natural accession
01-01-038	-75.7896	42.8020	New York	USA	Northeastern USA	Natural accession
01-01-042	-75.5857	42.5942	New York	USA	Northeastern USA	Natural accession
01-01-047	-76.5318	42.4108	New York	USA	Northeastern USA	Natural accession
01-01-051	-76.9151	42.3629	New York	USA	Northeastern USA	Natural accession

Table A2.1 (Continued)

01-01-054	-76.8456	42.3232	New York	USA	Northeastern USA	Natural accession
01-01-064	-77.0934	42.8549	New York	USA	Northeastern USA	Natural accession
01-01-078	-75.2680	43.2924	New York	USA	Northeastern USA	Natural accession
01-01-079	-75.3279	43.3219	New York	USA	Northeastern USA	Natural accession
01-01-082	-75.3321	43.3209	New York	USA	Northeastern USA	Natural accession
01-01-084	-76.1421	43.0308	New York	USA	Northeastern USA	Natural accession
01-01-094	-75.3640	42.5628	New York	USA	Northeastern USA	Natural accession
01-01-213	-78.5566	42.0905	New York	USA	Northeastern USA	Natural accession
01-03-187	-77.9212	41.7569	Pennsylvania	USA	Northeastern USA	Natural accession
01-03-198	-79.1182	41.0707	Pennsylvania	USA	Northeastern USA	Natural accession
01-03-208	-80.2690	41.2735	Pennsylvania	USA	Northeastern USA	Natural accession
01-03-212	Location not recorded		Pennsylvania	USA	Northeastern USA	Natural accession
01-07-251	-73.3553	41.9836	Connecticut	USA	Northeastern USA	Natural accession
01-08-257	-73.2539	42.8778	Vermont	USA	Northeastern USA	Natural accession
02-201-005	Location not recorded			UKR	Eastern European	Natural accession
03-01-005	-76.2608	43.0702	New York	USA	Northeastern USA	Natural accession
03-01-007	-76.2608	43.0700	New York	USA	Northeastern USA	Natural accession
03-01-013	-76.2615	43.0687	New York	USA	Northeastern USA	Natural accession
03-01-017	-76.2632	43.0693	New York	USA	Northeastern USA	Natural accession
03-01-019	-76.2637	43.0701	New York	USA	Northeastern USA	Natural accession
03-01-020	-76.2623	43.0702	New York	USA	Northeastern USA	Natural accession
03-01-022	-76.2591	43.0718	New York	USA	Northeastern USA	Natural accession
03-01-023	-76.2571	43.0721	New York	USA	Northeastern USA	Natural accession
03-01-024	-76.2555	43.0723	New York	USA	Northeastern USA	Natural accession
03-01-025	-76.2376	43.0698	New York	USA	Northeastern USA	Natural accession
03-01-026	-76.9955	42.8790	New York	USA	Northeastern USA	Natural accession
03-01-027	-76.9993	42.8790	New York	USA	Northeastern USA	Natural accession
03-01-036	-76.2295	43.8050	New York	USA	Northeastern USA	Natural accession

Table A2.1 (Continued)

05-01-001		-76.2608	43.0700	New York	USA	Northeastern USA	Natural accession
05-01-002		-76.2624	43.0702	New York	USA	Northeastern USA	Natural accession
05-01-003		-76.2531	43.0722	New York	USA	Northeastern USA	Natural accession
05-01-005		-76.2561	43.0707	New York	USA	Northeastern USA	Natural accession
06-01-003		-76.2569	43.0722	New York	USA	Northeastern USA	Natural accession
07-MBG-5095				Quebec	CAN	Quebec	Natural accession from Montreal Botanical Gardens
07-MBG-5096				Quebec	CAN	Quebec	Natural accession from Montreal Botanical Gardens
PMC910630				NY	USA	Northeastern USA	Natural accession from USADA-NRCS
2		Location not recorded					Natural accession from University of Toronto
Pur12		Location not recorded		Ontario	CAN	Ontario	Natural accession from University of Toronto
Pur34		Location not recorded		Ontario	CAN	Ontario	Natural accession from University of Toronto
04-202-055	'Denmark 601'				USA		Cultivar obtained from AgriGenesis
04-202-056	'Eugenei 239'				USA		Cultivar obtained from AgriGenesis
04-202-057	'Green Dick 609'				USA		Cultivar obtained from AgriGenesis
04-202-058	'Holland NZ 605'				USA		Cultivar obtained from AgriGenesis
04-202-059	'Irette PN 608'				USA		Cultivar obtained from AgriGenesis
04-202-060	'Links Dutch 382'				USA		Cultivar obtained from AgriGenesis
04-BN-044	'#187"				USA		Cultivar obtained from Blue Stem Nursery
04-BN-045	'Eugenii'				USA		Cultivar obtained from Blue Stem Nursery
04-BN-046	'Green Dicks'				USA		Cultivar obtained from Blue Stem Nursery
04-BN-047	'Lambertiana'				USA		Cultivar obtained from Blue Stem Nursery

Table A2.1 (Continued)

11-BN-012	'Irette'		USA		Cultivar obtained from Blue Stem Nursery
11-BN-013	'Nana'		USA		Cultivar obtained from Blue Stem Nursery
11-BN-014	'Pendula'		USA		Cultivar obtained from Blue Stem Nursery
05-OSU-041	'Dicky Meadows'		USA		Cultivar obtained from Ohio State University
05-OSU-063	'Lambertiana'		USA		Cultivar obtained from Ohio State University
9882-34	'Fish Creek'	NY	USA	Northeastern USA	Bred cultivar
9882-41	'Wolcott'	NY	USA	Northeastern USA	Bred cultivar
05X-293-047		NY	USA	Northeastern USA	Bred cultivar
	'Hotel'		CAN		Cultivar obtained from LandSaga Biogeographical Inc.
PU1		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU2		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU3		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU4		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU5		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU6		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU7		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU8		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU9		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU10		Brandenburg	DEU	European_Baltic Inland	Natural accession
PU11		Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU12		Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU13		Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU14		Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU15		Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU16		Brandenburg	DEU	European_Baltic-Oder River	Natural accession

Table A2.1 (Continued)

PU17			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU18			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU19			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU20			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU21			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU22			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU23			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU24			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU25			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU26			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU27			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU28			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU29			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU30			Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU31	53.0606	14.3190	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU32	53.0600	14.3191	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU33	53.0589	14.3182	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU34	53.0578	14.3168	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU35			Pomerania	POL	European_Baltic Sea	Natural accession
PU36			Pomerania	POL	European_Baltic Sea	Natural accession
PU37			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU38			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU39			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU40			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU41			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU42			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU43			Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU44			Bavaria	DEU	European_Bavarian Prealps	Natural accession

Table A2.1 (Continued)

PU45	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU46	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU47	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU48	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU49	Bavaria	DEU	European_Bavarian Alps	Natural accession
PU53	Bavaria	DEU	European_Bavarian Alps	Natural accession
PU54	Bavaria	DEU	European_Bavarian Alps	Natural accession
PU55	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU56	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU57	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU58	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU59	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU60	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU61	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU62	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU63	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU64	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU65	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU66	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU67	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU68	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU69	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU70	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU71	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU72	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU73	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU74	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU75	Bavaria	DEU	European_Bavarian Prealps	Natural accession

Table A2.1 (Continued)

PU76	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU77	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU78	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU79	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU80	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU81	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU82	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU83	Bavaria	DEU	European_Bavarian Prealps	Natural accession
PU84	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU85	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU86	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU88	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU89	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU90	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU91	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU92	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU93	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU94	Vorarlberg	AUT	European_Bavarian Prealps	Natural accession
PU95	Vorarlberg	AUT	European_Bavarian Prealps	Natural accession
PU96	Vorarlberg	AUT	European_Bavarian Prealps	Natural accession
PU97	Vorarlberg	AUT	European_Bavarian Prealps	Natural accession
PU98	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU99	Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession

Table A2.1 (Continued)

PU100			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU101			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU102			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU103			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU104			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU105			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU106			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU107			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU108			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU109			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU110			Baden- Württemberg	DEU	European_Bavarian Prealps	Natural accession
PU111	47.0217	12.3761	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU112	47.0218	12.3761	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU113	47.0270	12.3800	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU114	47.0216	12.3763	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU115	47.0209	12.3758	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU116	46.9950	12.6410	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU117	46.9961	12.6411	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU118	46.9977	12.6412	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU119	46.9978	12.6410	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU120	46.9966	12.6421	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU122	46.9157	12.3746	Tyrol	AUT	European_Austrian Central Alps	Natural accession
PU123	46.8805	12.1635	South Tyrol	ITA	European_Southern Limestone	Natural accession

Table A2.1 (Continued)

					Alps	
PU124	46.8805	12.1636	South Tyrol	ITA	European_Southern Limestone Alps	Natural accession
PU125	46.8821	12.1661	South Tyrol	ITA	European_Southern Limestone Alps	Natural accession
PU126	46.7312	12.2042	South Tyrol	ITA	European_Southern Limestone Alps	Natural accession
PU128	46.7685	12.5816	Tyrol	AUT	European_Southern Limestone Alps	Natural accession
PU129	46.7683	12.5824	Tyrol	AUT	European_Southern Limestone Alps	Natural accession
PU130	47.1875	12.4269	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU131	47.2131	12.4177	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU132	47.2269	12.4125	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU133	47.2270	12.4123	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU134	47.2579	12.4101	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU135	47.2851	12.7529	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU136	47.2850	12.7528	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU137	47.2985	12.8116	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU138	47.3004	12.8079	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU139	47.3014	12.8055	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU140	47.1743	12.1824	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU141	47.1755	12.1687	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU142	47.1752	12.1821	Salzburg	AUT	European_Austrian Central Alps	Natural accession
PU143	53.1432	14.3907	Pomerania West	POL	European_Baltic-Oder River	Natural accession
PU144	53.1431	14.3908	Pomerania West	POL	European_Baltic-Oder River	Natural accession
PU145	53.1434	14.3913	Pomerania West	POL	European_Baltic-Oder River	Natural accession
PU146	53.2541	14.4509	Pomerania	POL	European_Baltic-Oder River	Natural accession

Table A2.1 (Continued)

PU147	53.2543	14.4509	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU148	53.2542	14.4471	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU149	53.3751	14.5558	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU150	53.3747	14.5588	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU151	53.5516	14.5991	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU152	53.5512	14.5991	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU153	53.5511	14.5991	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU154	53.5512	14.5990	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU155	53.5507	14.5989	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU156	53.5505	14.5986	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU157	53.5504	14.5984	West Pomerania	POL	European_Baltic-Oder River	Natural accession
PU158	53.1526	14.3543	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU159	53.1116	14.3411	Brandenburg	DEU	European_Baltic-Oder River	Natural accession
PU160			West Pomerania	POL	European_Baltic-Oder River	Natural accession

Table A2.2 Assignment of genotypes to clusters based on affinity propagation optimized grouping.

CloneID	Population	AP Cluster
00-01-103	American	1
01-01-038	American	1
01-01-051	American	1
01-01-054	American	1
01-01-213	American	1
01-03-187	American	1
01-03-198	American	1
01-03-208	American	Exemplar 1
01-03-212	American	1
95002	American	1
95071	American	1
01-01-028	American	2
01-01-047	American	2
01-01-064	American	2
01-01-082	American	2
01-07-251	American	2
01-08-257	American	2
03-01-025	American	2
03-01-026	American	2
03-01-027	American	2
04-202-060	Cultivar	2
04-BN-044	Cultivar	2
04-BN-045	Cultivar	2
04-BN-047	Cultivar	2
95038	American	2
95042	American	Exemplar 2
95049	American	2
95057	American	2
95058	American	2
Hotel	American	2
PMC9106302	American	2
PU104	European-Bavarian-Prealps	2
Pur34	American	2
03-01-036	American	Exemplar 3
04-202-057	Cultivar	3
04-202-058	Cultivar	3
04-BN-046	Cultivar	3
05-OSU-041	Cultivar	3
05-OSU-063	Cultivar	3
07-MBG-5095	Cultivar	3

Table A2.2 (Continued)

07-MBG-5096	Cultivar	3
PU101	European-Bavarian-Prealps	3
PU102	European-Bavarian-Prealps	3
PU106	European-Bavarian-Prealps	3
PU107	European-Bavarian-Prealps	3
PU108	European-Bavarian-Prealps	3
PU109	European-Bavarian-Prealps	3
PU110	European-Bavarian-Prealps	3
PU49	European-Bavarian-Prealps	3
PU5	European-Baltic-Inland	3
PU6	European-Baltic-Inland	3
PU91	European-Bavarian-Prealps	3
04-202-056	Cultivar	4
04-202-059	Cultivar	4
11-BN-012	Cultivar	4
11-BN-013	Cultivar	4
11-BN-014	Cultivar	4
PU100	European-Bavarian-Prealps	4
PU103	European-Bavarian-Prealps	4
PU105	European-Bavarian-Prealps	4
PU111	European-Southern-Limestone-Alps	4
PU112	European-Southern-Limestone-Alps	4
PU113	European-Southern-Limestone-Alps	4
PU114	European-Southern-Limestone-Alps	4
PU115	European-Southern-Limestone-Alps	4
PU116	European-Southern-Limestone-Alps	4
PU118	European-Southern-Limestone-Alps	4
PU119	European-Southern-Limestone-Alps	4
PU120	European-Southern-Limestone-Alps	4
PU129	European-Southern-Limestone-Alps	4
PU122	European-Southern-Limestone-Alps	4
PU123	European-Southern-Limestone-Alps	4
PU124	European-Southern-Limestone-Alps	4
PU125	European-Southern-Limestone-Alps	4
PU126	European-Southern-Limestone-Alps	4
PU128	European-Southern-Limestone-Alps	4
PU130	European-Southern-Limestone-Alps	4
PU131	European-Southern-Limestone-Alps	4
PU132	European-Southern-Limestone-Alps	4
PU133	European-Southern-Limestone-Alps	4
PU134	European-Southern-Limestone-Alps	4

Table A2.2 (Continued)

PU135	European-Southern-Limestone-Alps	4
PU136	European-Southern-Limestone-Alps	4
PU137	European-Southern-Limestone-Alps	4
PU138	European-Southern-Limestone-Alps	4
PU139	European-Southern-Limestone-Alps	4
PU140	European-Southern-Limestone-Alps	4
PU141	European-Southern-Limestone-Alps	4
PU142	European-Southern-Limestone-Alps	4
PU155	European-Baltic-Oder-River	4
PU37	European-Bavarian-Prealps	4
PU38	European-Bavarian-Prealps	4
PU39	European-Bavarian-Prealps	4
PU40	European-Bavarian-Prealps	4
PU41	European-Bavarian-Prealps	4
PU42	European-Bavarian-Prealps	4
PU43	European-Bavarian-Prealps	4
PU44	European-Bavarian-Prealps	4
PU45	European-Bavarian-Prealps	4
PU46	European-Bavarian-Prealps	4
PU47	European-Bavarian-Prealps	4
PU48	European-Bavarian-Prealps	4
PU53	European-Bavarian-Prealps	4
PU54	European-Bavarian-Prealps	Exemplar 4
PU55	European-Bavarian-Prealps	4
PU56	European-Bavarian-Prealps	4
PU57	European-Bavarian-Prealps	4
PU58	European-Bavarian-Prealps	4
PU61	European-Bavarian-Prealps	4
PU62	European-Bavarian-Prealps	4
PU63	European-Bavarian-Prealps	4
PU64	European-Bavarian-Prealps	4
PU65	European-Bavarian-Prealps	4
PU66	European-Bavarian-Prealps	4
PU67	European-Bavarian-Prealps	4
PU68	European-Bavarian-Prealps	4
PU69	European-Bavarian-Prealps	4
PU70	European-Bavarian-Prealps	4
PU71	European-Bavarian-Prealps	4
PU72	European-Bavarian-Prealps	4
PU73	European-Bavarian-Prealps	4
PU74	European-Bavarian-Prealps	4

Table A2.2 (Continued)

PU75	European-Bavarian-Prealps	4
PU76	European-Bavarian-Prealps	4
PU77	European-Bavarian-Prealps	4
PU78	European-Bavarian-Prealps	4
PU79	European-Bavarian-Prealps	4
PU80	European-Bavarian-Prealps	4
PU81	European-Bavarian-Prealps	4
PU82	European-Bavarian-Prealps	4
PU83	European-Bavarian-Prealps	4
PU84	European-Bavarian-Prealps	4
PU85	European-Bavarian-Prealps	4
PU86	European-Bavarian-Prealps	4
PU88	European-Bavarian-Prealps	4
PU89	European-Bavarian-Prealps	4
PU90	European-Bavarian-Prealps	4
PU92	European-Bavarian-Prealps	4
PU93	European-Bavarian-Prealps	4
PU94	European-Bavarian-Prealps	4
PU95	European-Bavarian-Prealps	4
PU96	European-Bavarian-Prealps	4
PU97	European-Bavarian-Prealps	4
PU98	European-Bavarian-Prealps	4
PU99	European-Bavarian-Prealps	4
Pur12	American	4
00-01-106	American	5
02-201-005	American	5
04-202-055	Cultivar	5
PU10	European-Baltic-Inland	5
PU117	European-Southern-Limestone-Alps	5
PU11	European-Baltic-Oder-River	5
PU12	European-Baltic-Oder-River	5
PU13	European-Baltic-Oder-River	5
PU160	European-Baltic-Oder-River	5
PU143	European-Baltic-Oder-River	5
PU144	European-Baltic-Oder-River	5
PU145	European-Baltic-Oder-River	5
PU146	European-Baltic-Oder-River	5
PU147	European-Baltic-Oder-River	5
PU148	European-Baltic-Oder-River	5
PU149	European-Baltic-Oder-River	5
PU14	European-Baltic-Oder-River	5

Table A2.2 (Continued)

PU150	European-Baltic-Oder-River	5
PU151	European-Baltic-Oder-River	5
PU152	European-Baltic-Oder-River	5
PU153	European-Baltic-Oder-River	5
PU154	European-Baltic-Oder-River	5
PU156	European-Baltic-Oder-River	5
PU157	European-Baltic-Oder-River	5
PU158	European-Baltic-Oder-River	5
PU159	European-Baltic-Oder-River	5
PU15	European-Baltic-Oder-River	5
PU16	European-Baltic-Oder-River	5
PU17	European-Baltic-Oder-River	5
PU18	European-Baltic-Oder-River	5
PU19	European-Baltic-Oder-River	5
PU1	European-Baltic-Inland	5
PU20	European-Baltic-Oder-River	5
PU21	European-Baltic-Oder-River	5
PU22	European-Baltic-Oder-River	5
PU23	European-Baltic-Oder-River	5
PU24	European-Baltic-Oder-River	5
PU25	European-Baltic-Oder-River	5
PU26	European-Baltic-Oder-River	5
PU27	European-Baltic-Oder-River	5
PU28	European-Baltic-Oder-River	5
PU29	European-Baltic-Oder-River	5
PU2	European-Baltic-Inland	exemplar 5
PU30	European-Baltic-Oder-River	5
PU31	European-Baltic-Oder-River	5
PU32	European-Baltic-Oder-River	5
PU33	European-Baltic-Oder-River	5
PU34	European-Baltic-Oder-River	5
PU35	European-Baltic-Sea	5
PU36	European-Baltic-Sea	5
PU3	European-Baltic-Inland	5
PU4	European-Baltic-Inland	5
PU59	European-Bavarian-Prealps	5
PU60	European-Bavarian-Prealps	5
PU7	European-Baltic-Inland	5
PU8	European-Baltic-Inland	5
PU9	European-Baltic-Inland	5
00-01-001	American	6

Table A2.2 (Continued)

00-01-003	American	6
00-01-009	American	6
00-01-011	American	6
00-01-014	American	6
00-01-034	American	6
00-01-086	American	6
00-01-088	American	6
00-01-091	American	6
00-01-094	American	6
00-01-095	American	6
00-01-101	American	6
00-01-104	American	6
00-01-105	American	6
01-01-001	American	6
01-01-032	American	6
01-01-034	American	exemplar 6
01-01-036	American	6
01-01-084	American	6
03-01-005	American	6
03-01-007	American	6
03-01-013	American	6
03-01-017	American	6
03-01-023	American	6
03-01-024	American	6
05X-293-047	Cultivar	6
05-01-001	American	6
05-01-002	American	6
05-01-003	American	6
05-01-005	American	6
94001	American	6
94002	American	6
94005	American	6
94009	American	6
94011	American	6
94012	American	6
94013	American	6
94014	American	6
95026	American	6
9882-34	Cultivar	6
9882-41	Cultivar	6
00-01-004	American	7

Table A2.2 (Continued)

00-01-085	American	7
00-01-089	American	7
00-01-098	American	7
00-01-102	American	7
00-22-002	American	7
01-01-029	American	7
01-01-030	American	7
01-01-031	American	7
01-01-042	American	7
01-01-078	American	7
01-01-079	American	7
01-01-094	American	7
03-01-019	American	exemplar 7
03-01-020	American	7
03-01-022	American	7
06-01-003	American	7
94003	American	7
94004	American	7
94006	American	7
94015	American	7
95001	American	7
95005	American	7

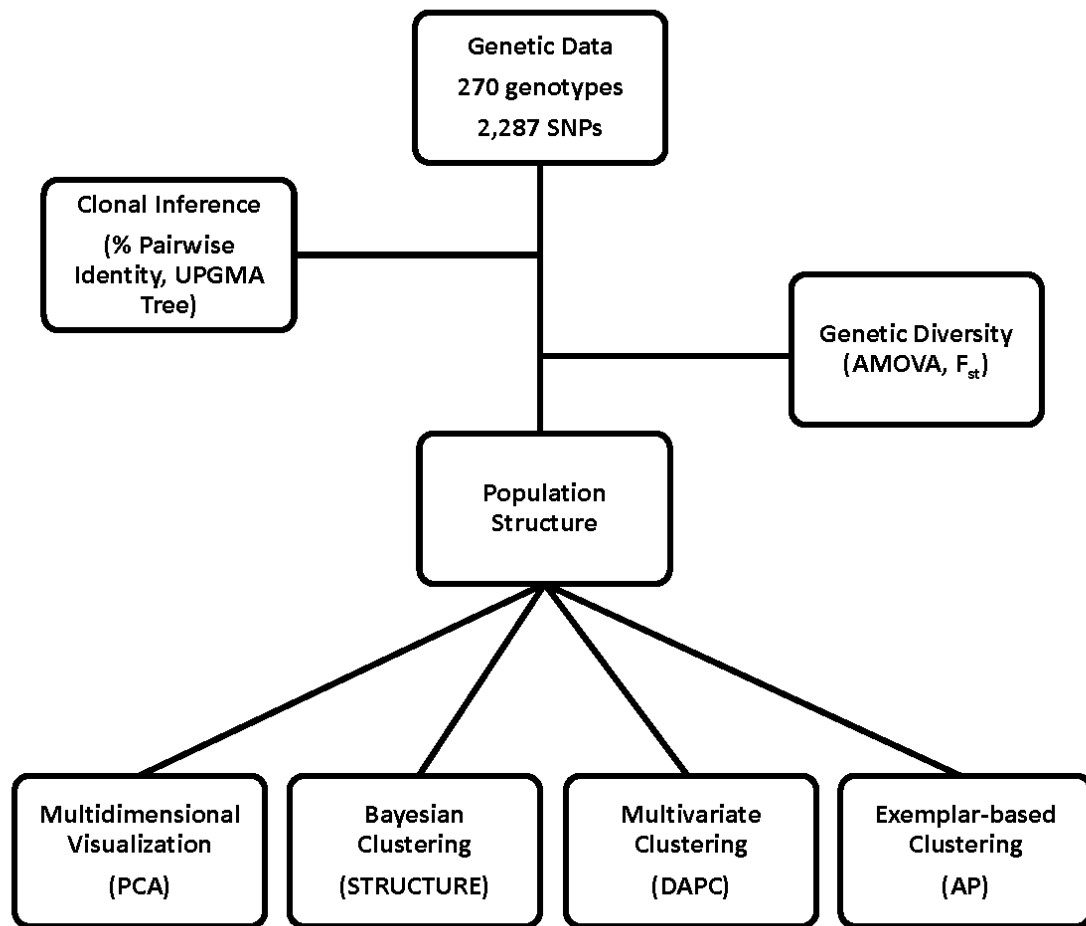
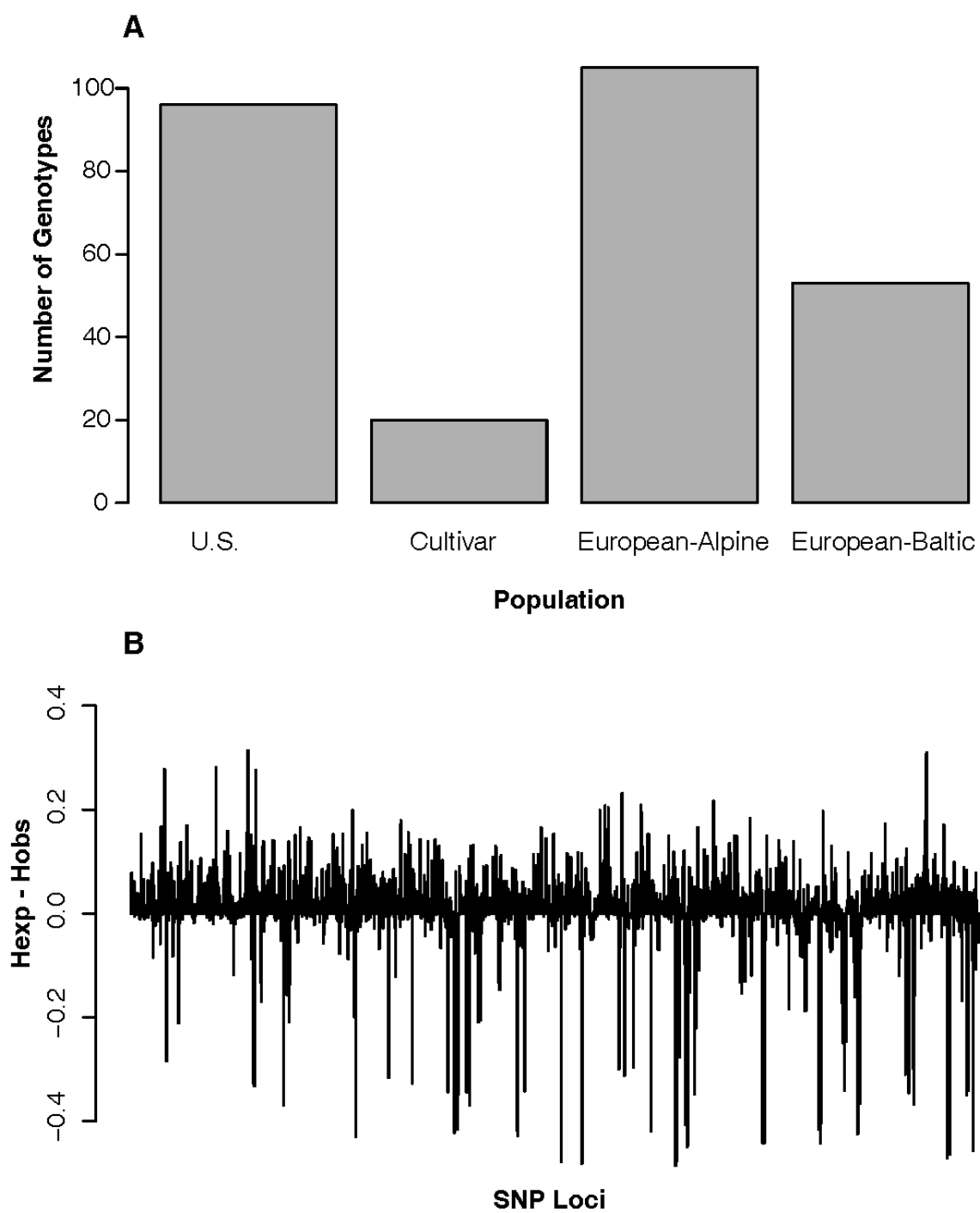


Figure A2.1 Diagram outlining data analysis used in this study



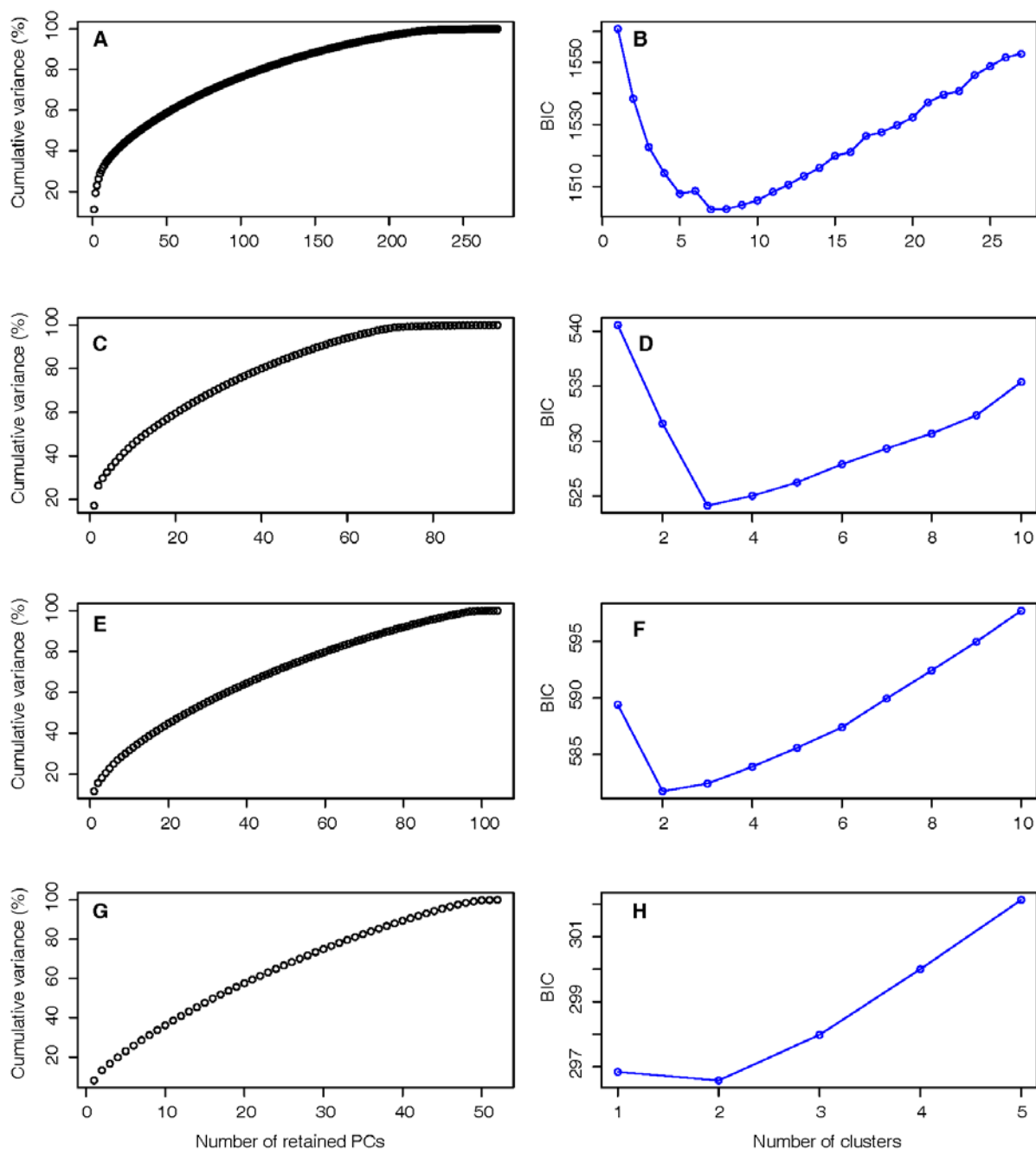


Figure A2.3 *K*-means hierarchical clustering for 267 genotypes. Percentage of variance explained for the number of retained principal components (PCs) and the Bayesian information criterion (BIC) indicating the optimum number of inferred clusters for each population for A-B) all populations C-D) US accessions and cultivars E-F) European Alpine population and G-H) European Baltic population.

APPENDIX TO CHAPTER 3

Table A3.1 Clone ID, sex, and source information for 110 genotypes in the diverse *S. purpurea* collection

Clone ID	Epithet	Sex	Species	Latitude	Longitude	City/Town	State	Country	Notes
94001		M	<i>S. purpurea</i>	N 43 13.000	W 75 38	Rome	NY	US	Natural accession
94002		M	<i>S. purpurea</i>	N 43 13.001	W 75 38	Rome	NY	US	Natural accession
94003		H	<i>S. purpurea</i>	N 43 13.002	W 75 38	Rome	NY	US	Natural accession
94004		F	<i>S. purpurea</i>	N 43 13.003	W 75 38	Rome	NY	US	Natural accession
94005		F	<i>S. purpurea</i>	N 43 13.004	W 75 38	Rome	NY	US	Natural accession
94006		F	<i>S. purpurea</i>	N 43 13.005	W 75 38	Rome	NY	US	Natural accession
94009		M	<i>S. purpurea</i>	N 42 52.002	W 75 51	Cazenovia	NY	US	Natural accession
94011		F	<i>S. purpurea</i>	N 42 52.004	W 75 51	Cazenovia	NY	US	Natural accession
94012		F	<i>S. purpurea</i>	N 42 52.005	W 75 51	Cazenovia	NY	US	Natural accession
94013		F	<i>S. purpurea</i>	N 42 52.006	W 75 51	Cazenovia	NY	US	Natural accession
94014		F	<i>S. purpurea</i>	N 42 52.007	W 75 51	Cazenovia	NY	US	Natural accession
94015		F	<i>S. purpurea</i>	N 42 52.008	W 75 51	Cazenovia	NY	US	Natural accession

Table A3.1 (Continued)

95001	F	<i>S. purpurea</i>	N 42 51	W 75 49	New Woodstock	NY	US	Natural accession
95002	M	<i>S. purpurea</i>	N 42 51.035	W 75 46	Nelson	NY	US	Natural accession
95005	F	<i>S. purpurea</i>	N 42 50	W 75 47.5	Erieville	NY	US	Natural accession
95026	F	<i>S. purpurea</i>	N 41 46.3	W 73 33	Wassaic	NY	US	Natural accession
95038	M	<i>S. purpurea</i>	N 42 49	W 78 8	Wyoming	NY	US	Natural accession
95042	M	<i>S. purpurea</i>	N 42 49	W 78 8	Wyoming	NY	US	Natural accession
95049	M	<i>S. purpurea</i>	N 42 22	W 79 14	Cassadaga	NY	US	Natural accession
95057	M	<i>S. purpurea</i>	N 42 18	W 78 15	Caneadea	NY	US	Natural accession
95058	M	<i>S. purpurea</i>	N 42 18	W 78 15	Caneadea	NY	US	Natural accession
95071	M	<i>S. purpurea</i>	N 42 20	W 76 50	Montour Falls	NY	US	Natural accession
00-01-001	M	<i>S. purpurea</i>	N 43 21.57	W 75 32.6	Lee	NY	US	Natural accession
00-01-003	F	<i>S. purpurea</i>	N 43 21.585	W 75 32.695	Lee	NY	US	Natural accession

Table A3.1 (Continued)

00-01-004	F	<i>S. purpurea</i>	N 43 21.595	W 75 32.758	Lee	NY	US	Natural accession
00-01-009	F	<i>S. purpurea</i>	N 43 23.986	W 75 32.43	Ava	NY	US	Natural accession
00-01-011	F	<i>S. purpurea</i>	N 43 23.994	W 75 32.43	Ava	NY	US	Natural accession
00-01-014	M	<i>S. purpurea</i>	N 43 10.06	W 75 40.77	Durhamville	NY	US	Natural accession
00-01-034	F	<i>S. purpurea</i>	N 42 47.352	W 76 11.434	Tully	NY	US	Natural accession
00-01-085	M	<i>S. purpurea</i>	N 42 53.707	W 76 2.432	LaFayette	NY	US	Natural accession
00-01-086	F	<i>S. purpurea</i>	N 42 53.639	W 76 2.237	Pompey	NY	US	Natural accession
00-01-088	F	<i>S. purpurea</i>	N 42 49.493	W 76 0.286	Fabius	NY	US	Natural accession
00-01-089	M	<i>S. purpurea</i>	N 42 50.325	W 75 57.091	Fabius	NY	US	Natural accession
00-01-091	M	<i>S. purpurea</i>	N 42 44.682	W 75 36.75	Earlville	NY	US	Natural accession
00-01-094	M	<i>S. purpurea</i>	N 42 43.997	W 75 35.04	Earlville	NY	US	Natural accession
00-01-095	M	<i>S. purpurea</i>	N 42 33.185	W 75 26.586	New Berlin	NY	US	Natural accession

Table A3.1 (Continued)

00-01-098	F	<i>S. purpurea</i>	N 42 27.983	W 75 25.426	Mt. Upton	NY	US	Natural accession
00-01-101	F	<i>S. purpurea</i>	N 42 54.168	W 75 40.253	Morrisville	NY	US	Natural accession
00-01-102	M	<i>S. purpurea</i>	N 42 52.043	W 75 40.82	Eaton	NY	US	Natural accession
00-01-103	F	<i>S. purpurea</i>	N 43 48.058	W 75 38.653	Lowville	NY	US	Natural accession
00-01-104	F	<i>S. purpurea</i>	N 43 40.538	W 75 26.473	Turin	NY	US	Natural accession
00-01-105	M	<i>S. purpurea</i>	N 43 34.208	W 75 25.167	Constableville	NY	US	Natural accession
00-01-106	F	<i>S. purpurea</i>	N 43 21.942	W 75 29.598	Lee Center	NY	US	Natural accession
00-22-002	H	<i>S. purpurea</i>	Location not recorded			NC	US	Natural accession
01-01-001	M	<i>S. purpurea</i>	N 43 2.618	W 76 3.021	Fayetteville	NY	US	Natural accession
01-01-028	F	<i>S. purpurea</i>	N 42 45.016	W 76 7.893	Tully	NY	US	Natural accession
01-01-029	M	<i>S. purpurea</i>	N 42 54.388	W 75 54.572	Cazenovia	NY	US	Natural accession
01-01-030	F	<i>S. purpurea</i>	N 42 54.365	W 75 54.623	Cazenovia	NY	US	Natural accession

Table A3.1 (Continued)

01-01-031	M	<i>S. purpurea</i>	N 42 54.386	W 75 54.48	Cazenovia	NY	US	Natural accession
01-01-032	F	<i>S. purpurea</i>	N 42 53.42	W 75 49.923	Cazenovia	NY	US	Natural accession
01-01-034	F	<i>S. purpurea</i>	N 42 53.183	W 75 49.52	Erieville	NY	US	Natural accession
01-01-036	F	<i>S. purpurea</i>	N 42 51.235	W 75 49.111	New Woodstock	NY	US	Natural accession
01-01-038	F	<i>S. purpurea</i>	N 42 48.12	W 75 47.377	Georgetown	NY	US	Natural accession
01-01-042	M	<i>S. purpurea</i>	N 42 35.654	W 75 35.141	South Plymouth	NY	US	Natural accession
01-01-047	M	<i>S. purpurea</i>	N 42 24.649	W 76 31.905	Ithaca	NY	US	Natural accession
01-01-051	M	<i>S. purpurea</i>	N 42 21.775	W 76 54.905	Watkins Glen	NY	US	Natural accession
01-01-054	M	<i>S. purpurea</i>	N 42 19.389	W 76 50.737	Millport	NY	US	Natural accession
01-01-064	M	<i>S. purpurea</i>	N 42 51.296	W 77 5.606	Stanley	NY	US	Natural accession
01-01-078	M	<i>S. purpurea</i>	N 43 17.546	W 75 16.081	Holland Patent	NY	US	Natural accession
01-01-079	M	<i>S. purpurea</i>	N 43 19.315	W 75 19.673	Westernville	NY	US	Natural accession

Table A3.1 (Continued)

01-01-082	F	<i>S. purpurea</i>	N 43 19.255	W 75 19.924	Westernville	NY	US	Natural accession
01-01-084	F	<i>S. purpurea</i>	N 43 1.848	W 76 8.528	Syracuse	NY	US	Natural accession
01-01-094	F	<i>S. purpurea</i>	N 42 33.767	W 75 21.842	New Berlin	NY	US	Natural accession
01-01-213	M	<i>S. purpurea</i>	N 42 5.43	W 78 33.394	Allegany	NY	US	Natural accession
01-03-187	M	<i>S. purpurea</i>	N 41 45.414	W 77 55.271	Coudersport	PA	US	Natural accession
01-03-198	M	<i>S. purpurea</i>	N 41 4.24	W 79 7.092	Brookville	PA	US	Natural accession
01-03-208	M	<i>S. purpurea</i>	N 41 16.409	W 80 16.14	Mercer	PA	US	Natural accession
01-03-212	M	<i>S. purpurea</i>	Location not recorded			PA	US	Natural accession
01-07-251	F	<i>S. purpurea</i>	N 41 59.013	W 73 21.315	Canaan	CT	US	Natural accession
01-08-257	M	<i>S. purpurea</i>	N 42 52.667	W 73 15.232	Bennington	VT	US	Natural accession
02-201-005	M	<i>S. purpurea</i>	Location not recorded				UKR	Natural accession
03-01-005	F	<i>S. purpurea</i>	N 43 04.214	W 76 15.649	Solvay	NY	US	Natural accession

Table A3.1 (Continued)

03-01-007	F	<i>S. purpurea</i>	N 43 04.199	W 76 15.649	Solvay	NY	US	Natural accession
03-01-013	F	<i>S. purpurea</i>	N 43 04.119	W 76 15.689	Solvay	NY	US	Natural accession
03-01-017	F	<i>S. purpurea</i>	N 43 04.157	W 76 15.791	Solvay	NY	US	Natural accession
03-01-019	F	<i>S. purpurea</i>	N 43 04.207	W 76 15.819	Solvay	NY	US	Natural accession
03-01-020	F	<i>S. purpurea</i>	N 43 04.211	W 76 15.740	Syracuse	NY	US	Natural accession
03-01-022	M	<i>S. purpurea</i>	N 43 04.309	W 76 15.548	Syracuse	NY	US	Natural accession
03-01-023	F	<i>S. purpurea</i>	N 43 04.324	W 76 15.425	Syracuse	NY	US	Natural accession
03-01-024	M	<i>S. purpurea</i>	N 43 04.337	W 76 15.330	Syracuse	NY	US	Natural accession
03-01-025	M	<i>S. purpurea</i>	N 43 04.187	W 76 14.256	Syracuse	NY	US	Natural accession
03-01-026	M	<i>S. purpurea</i>	N 42 52.737	W 76 59.732	Geneva	NY	US	Natural accession
03-01-027	M	<i>S. purpurea</i>	N 42 52.737	W 76 59.957	Geneva	NY	US	Natural accession
03-01-036	F	<i>S. purpurea</i>	N 43 48.299	W 76 13.767	Henderson	NY	US	Natural accession

Table A3.1 (Continued)

04-202-055	'Denmark 601'	M	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-202-056	'Eugenei 239'	M	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-202-057	'Green Dick 609'	F	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-202-058	'Holland NZ 605'	M	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-202-059	'Irette PN 608'	M	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-202-060	'Links Dutch 382'	M	<i>S. purpurea</i>							Cultivar obtained from AgriGenesis
04-BN-044	'#187"	M	<i>S. purpurea</i>							Cultivar obtained from Blue Stem Nursery
04-BN-045	'Eugenii'	M	<i>S. purpurea</i>							Cultivar obtained from Blue Stem Nursery
04-BN-046	'Green Dicks'	F	<i>S. purpurea</i>							Cultivar obtained from Blue Stem Nursery
04-BN-047	'Lambertiana'	M	<i>S. purpurea</i>							Cultivar obtained from Blue Stem Nursery
05-01-001		M	<i>S. purpurea</i>	N 43 04.202	W 76 15.645	Syracuse	NY	US		Natural accession
05-01-002		F	<i>S. purpurea</i>	N 43 04.212	W 76 15.743	Syracuse	NY	US		Natural accession

Table A3.1 (Continued)

05-01-003		F	<i>S. purpurea</i>	N 43 04.33	W 76 15.187	Syracuse	NY	US	Natural accession
05-01-005		F	<i>S. purpurea</i>	N 43 04.240	W 76 15.368	Syracuse	NY	US	Natural accession
05-OSU-041	'Dicky Meadows'	F	<i>S. purpurea</i>					US	Cultivar obtained from Ohio State University
05-OSU-063	'Lambertiana'	F	<i>S. purpurea</i>					US	Cultivar obtained from Ohio State University
06-01-003		H	<i>S. purpurea</i>	N 43 04.333	W 76 15.413	Syracuse	NY	US	Natural accession
07-MBG-5095		F	<i>S. purpurea</i>					CAN	Natural accession obtained from Montreal Botanical Gardens
07-MBG-5096		F	<i>S. purpurea</i>					CAN	Natural accession obtained from Montreal Botanical Gardens
9882-34	'Fish Creek'	M	<i>S. purpurea</i>					US	Bred cultivar
9882-41	'Wolcott'	F	<i>S. purpurea</i>					US	Bred cultivar
	'Hotel'	M	<i>S. purpurea</i>					CAN	Cultivar obtained from LandSaga Biogeographical Inc.
PMC9106302		F	<i>S. purpurea</i>	Location not recorded				US	Natural accession obtained from USDA- NRCS

Table A3.1 (Continued)

Pur12	M	<i>S. purpurea</i>	Location not recorded	CAN	Natural accession obtained from University of Toronto
Pur34	M	<i>S. purpurea</i>	Location not recorded	CAN	Natural accession obtained from University of Toronto
05X-293-047	M	<i>S. purpurea</i>		US	Bred cultivar

Table A3.2 Experimental site characteristics for all trial locations

Site Characteristics ^a	Geneva, NY		Portland, NY	Morgantown, WV
	F ₁ & F ₂ Trial	Diverse Collection	Diverse Collection	Diverse Collection
Latitude	42°52'47"N	42°52'11"N	42°22'26"N	39°39'31"N
Longitude	77°00'55"W	77°03'10"W	79°29'11"W	79°54'19"W
Elevation (m)	184	234	228	365
Soil Type	Odessa silt and lima loam	Lima Loam	Gravelly loam	Dormont and Guernsey silt loam
Nitrate (mg kg ⁻¹)	-	1.3 ± 0.7	10.2 ± 1.7	4.3 ± 2.1
pH	-	6.8 ± 0.0	5.1 ± 0.0	6.3 ± 0.2
Organic (%)	-	2.5 ± 0.0	3.8 ± 0.1	4.4 ± 0.3
2012 GDD ^b		3041	2990	3819
2013 GDD		2731	2654	3548
2014 GDD		2678	2499	3576
2015 GDD		2730	2835	2787
2012 Precipitation (May – August; cm)		24.53	30.40	29.31
2013 Precipitation (May – August; cm)		46.38	44.22	52.17
2014 Precipitation (May – August; cm)		44.73	42.09	49.56
2015 Precipitation (May – August; cm)		40.03	13.21	33.17

^aWeather data for Geneva and Portland, NY were obtained from Cornell University's Network for Environment and Weather Applications database. Weather data for Morgantown, WV was collected from the National Oceanic and Atmospheric Administration website.

Means ± standard error are shown for nitrate, pH , and organic matter

^bGDD; growing degree days

Table A3.3 Summary of phenotypic traits from the diverse *S. purpurea* collection

Trait^a	Mean (\pmSE)	Min. – Max.
2013		
SDIA	7.61 (0.03)	1.15-15.55
SA	9.31 (6.88)	0.07-46.60
HT	1.93 (0.53)	0.11-3.54
IL	13.63 (4.93)	3.50-57.00
SNo	18.2 (10)	2.2-54.1
LFL	6.92 (1.67)	1.25-15.87
LFW	1.74 (1.29)	0.46-19.64
LFA	8.91 (4.43)	0.58-51.72
LFP	24.22 (26.22)	2.82-239.36
LFWT	0.08 (0.01)	0.01 -0.73
SLA	129.95 (93.43)	6.66-171.56
AugSPAD	46.52 (9.37)	17.13-75.00
SeptSPAD	42.20 (9.01)	17.10-77.30
g_s	596.41 (195.76)	45.50-1178.60
HEMI	17.58 (0.89)	14.41 -21.06
CELL	37.77 (3.26)	25.92 - 47.30
LIG	28.86 (2.07)	24.37-36.98
ASH	2.13 (0.59)	0.73-4.33
SPGR	0.50 (0.07)	0.23-0.79
2014		
SDIA	10.56 (0.06)	3.00-28.78
SA	21.24 (13.45)	0.07-84.38
HT	3.12 (0.67)	0.40-4.88
IL	13.45 (5.60)	4.55-32.00
SNo	23.6 (12)	1.2-73.4
CDIA	31.93 (19.76)	1.00-448.20
CDFOR	47.33 (14.34)	3.89-88.12
LFL	6.64 (1.86)	0.90-16.17
LFW	1.80 (1.70)	0.69-25.73
LFA	8.75 (3.97)	0.99-37.47
LFP	17.59 (8.72)	2.50-68.89
LFWT	0.11 (0.01)	0.02-0.47
SLA	90.78 (22.65)	18.83-419.34
YLD	9.23 (0.14)	0.00-55.58
VPH	109.6 (3.8)	102.0-118.0
FPH	91.2 (14.9)	57.0-115.0
AugSPAD	46.16 (5.74)	28.77-69.90
SeptSPAD	43.51 (7.80)	16.03-91
g_s	484.31 (173.29)	64.50-990.60
HEMI	17.66 (0.74)	15.87-21.77
CELL	41.63 (1.71)	33.27-46.01

Table A3.3 (Continued)

LIG	27.40 (1.16)	24.17-32.38
ASH	1.56 (0.38)	0.76-3.57
SPGR	0.49 (0.04)	0.32-0.78

^aPhenotypic traits measured in years 2013 (-13) and 2014 (-14). See Materials and Methods for trait definitions.

Table A3.4 Parameter estimates and significance values for multiple linear regression predictors of second year yield

Variable^a	Estimate	<i>t</i>	<i>P</i>-value	95% Confidence	
Intercept (β_0)	-2.56	-12.24	<.0001	-2.97	-2.15
SA-13 (β_1)	0.06	8.96	<.0001	0.05	0.08
SA-14 (β_2)	0.92	20.95	<.0001	0.83	1.00
HT-14 (β_3)	0.05	13.59	<.0001	0.04	0.06
AugSPAD-14 (β_4)	0.01	3.48	0.0005	0.01	0.02

^aSignificant predictor variables measured in 2013 (-13) and 2014 (-14).

Figure A3.1 A) Matrix of all pair-wise comparisons between traits by location within each year for Geneva, NY 2013. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

A

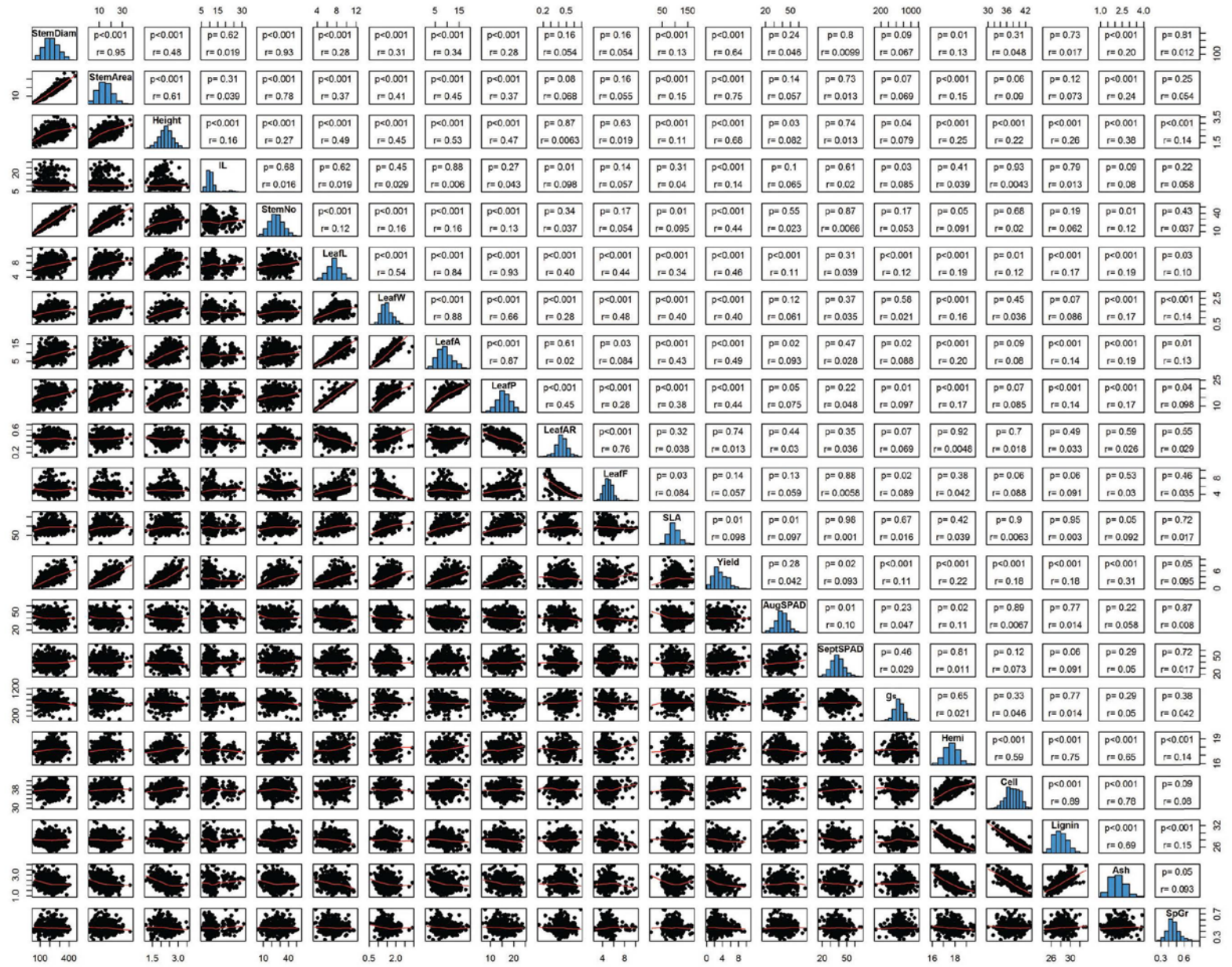


Figure A3.1 B) Matrix of all pair-wise comparisons between traits by location within each year for Portland, NY 2013. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

B

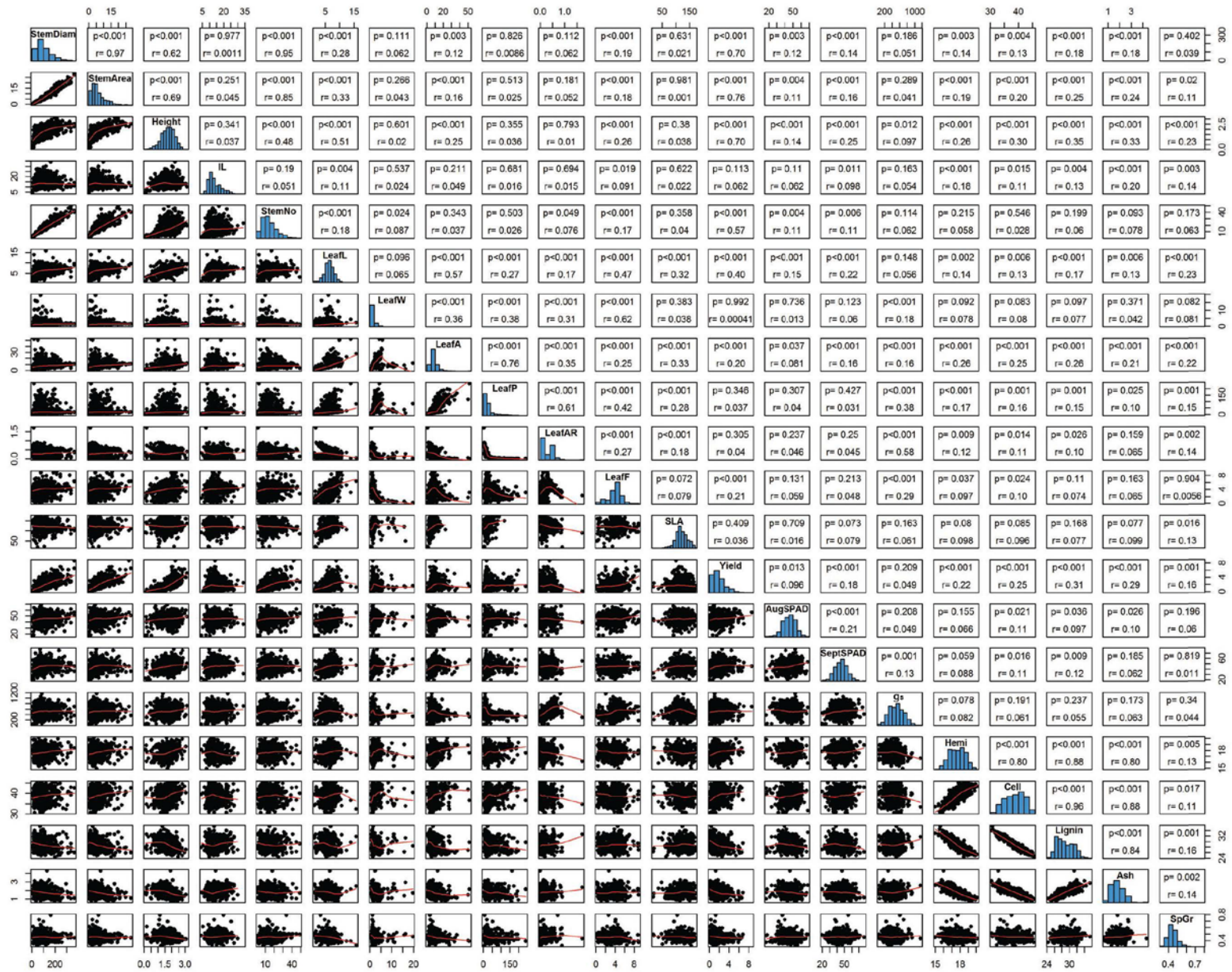


Figure A3.1 C) Matrix of all pair-wise comparisons between traits by location within each year for Morgantown, WV 2013. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

C

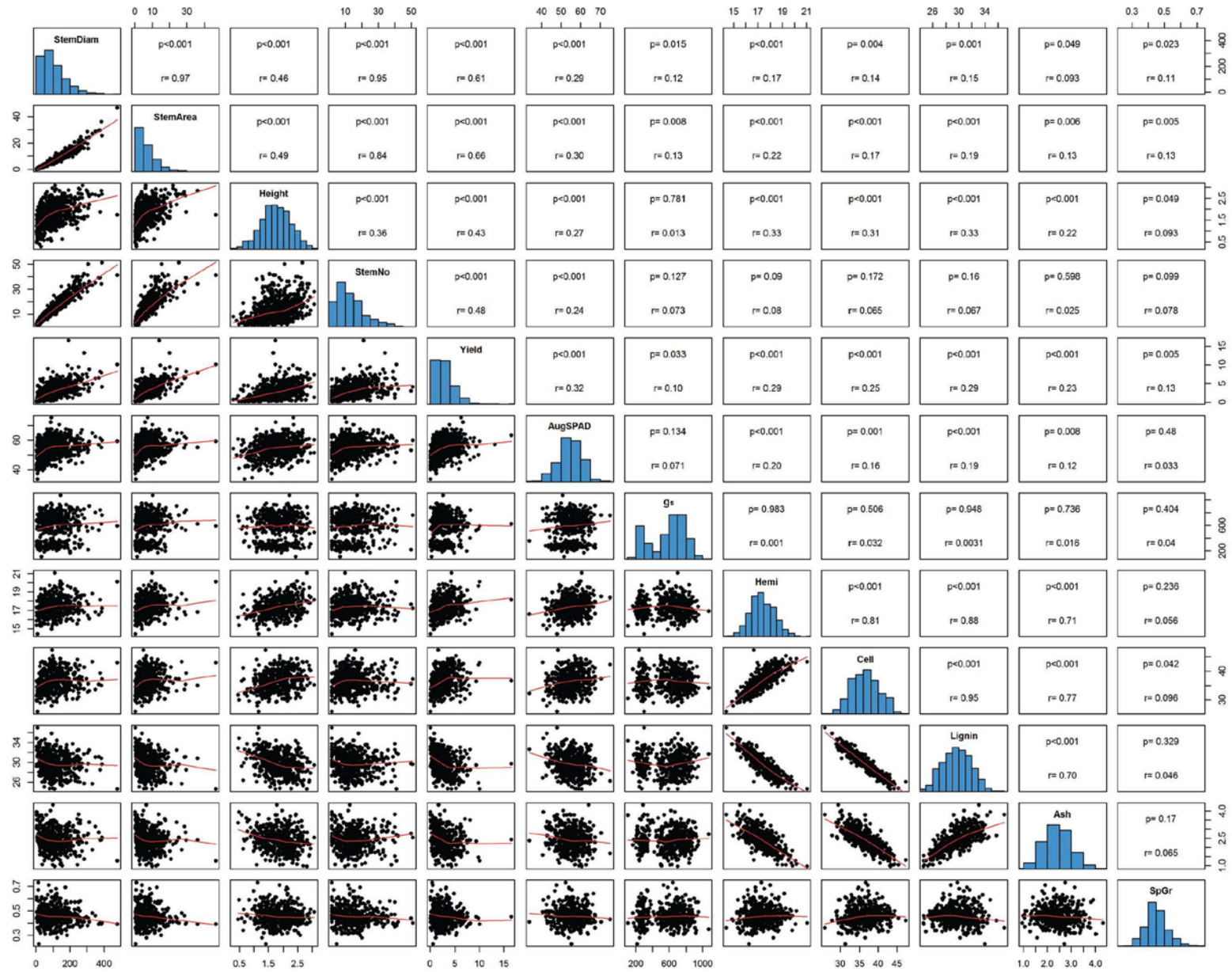


Figure A3.1 D) Matrix of all pair-wise comparisons between traits by location within each year for Geneva, NY 2014. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

D

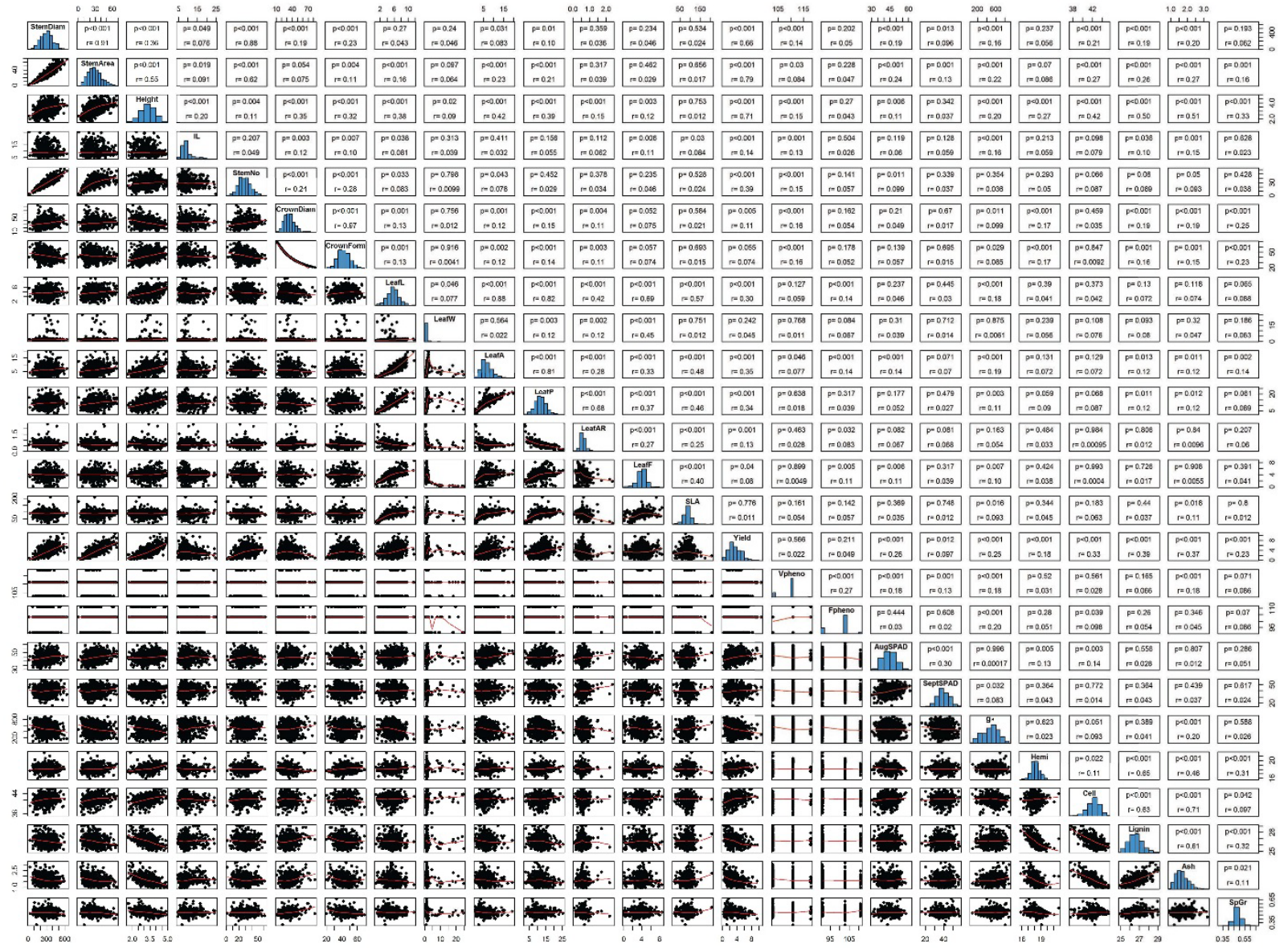


Figure A3.1 E) Matrix of all pair-wise comparisons between traits by location within each year for Portland, NY 2014. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

E

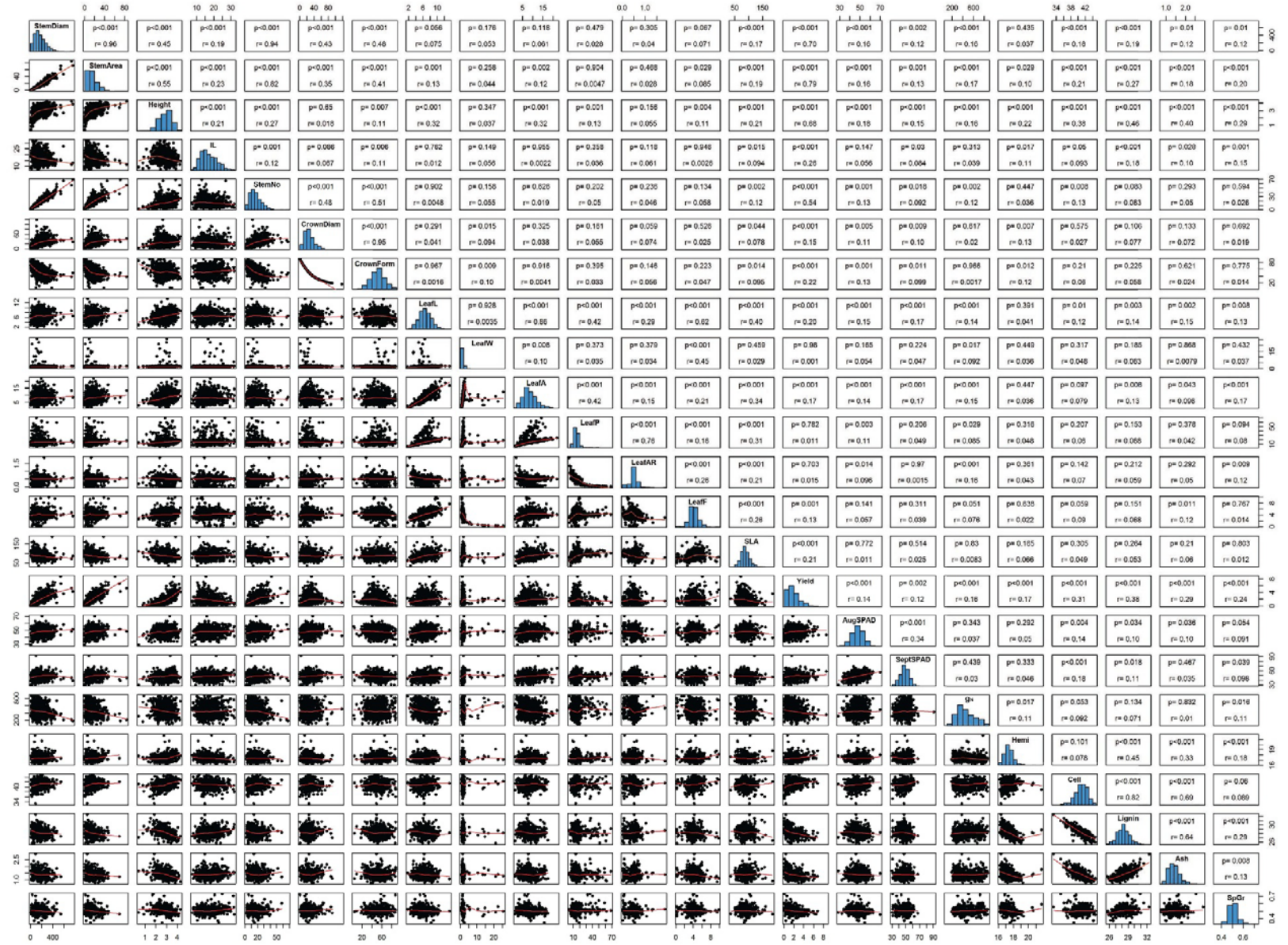


Figure A3.1 F) Matrix of all pair-wise comparisons between traits by location within each year for Morgantown, WV 2014. The lower diagonal shows a scatter plot matrix with a LOESS smooth curve fitting, the main diagonal is a histogram showing the distribution of each trait, and the upper diagonal indicating the Pearson correlation coefficient (r) and P -value for each comparison.

F

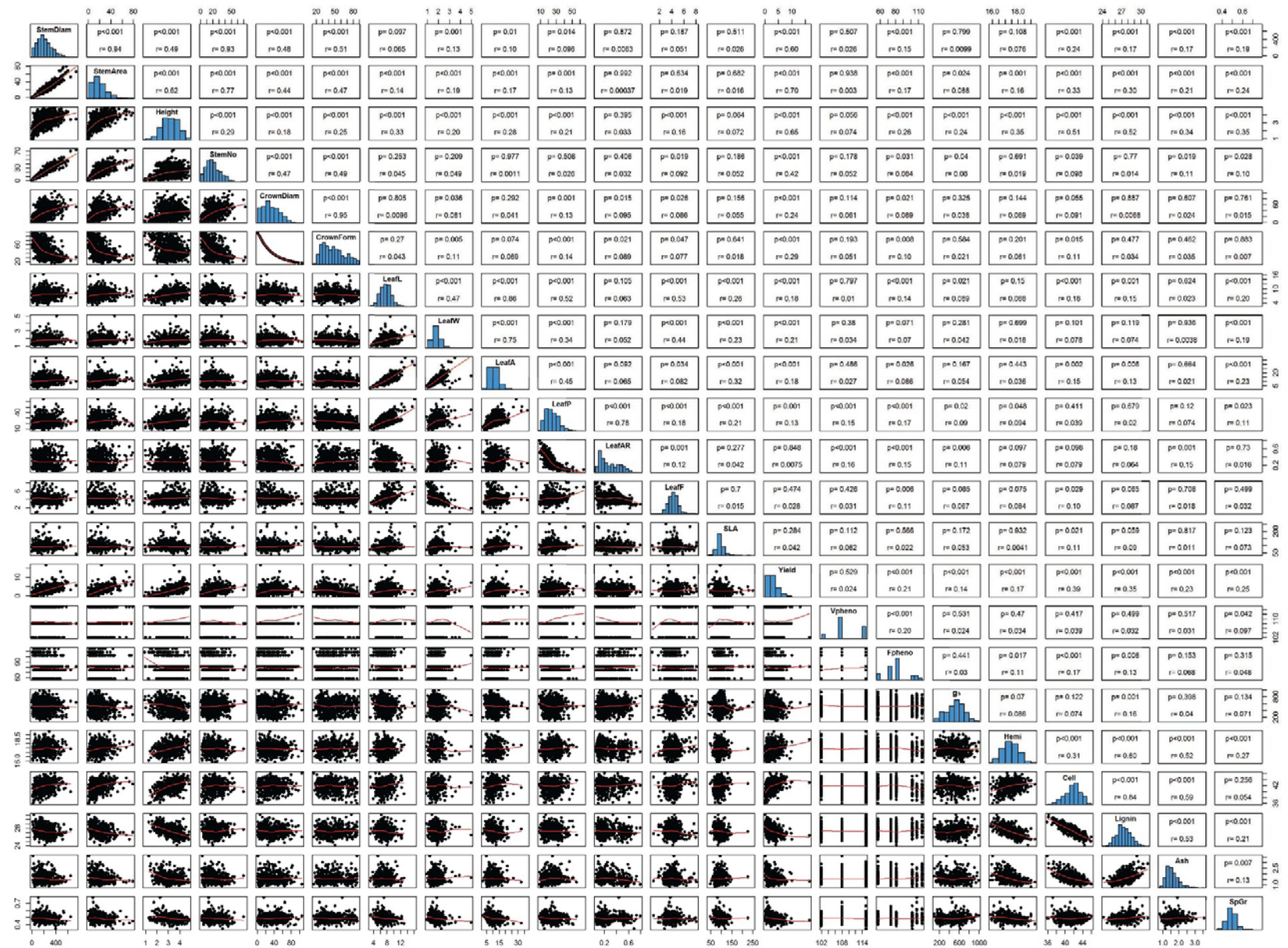


Figure A3.2 Matrix of all pair-wise comparisons between traits measured in 2015 for the *S. purpurea* F₁ family

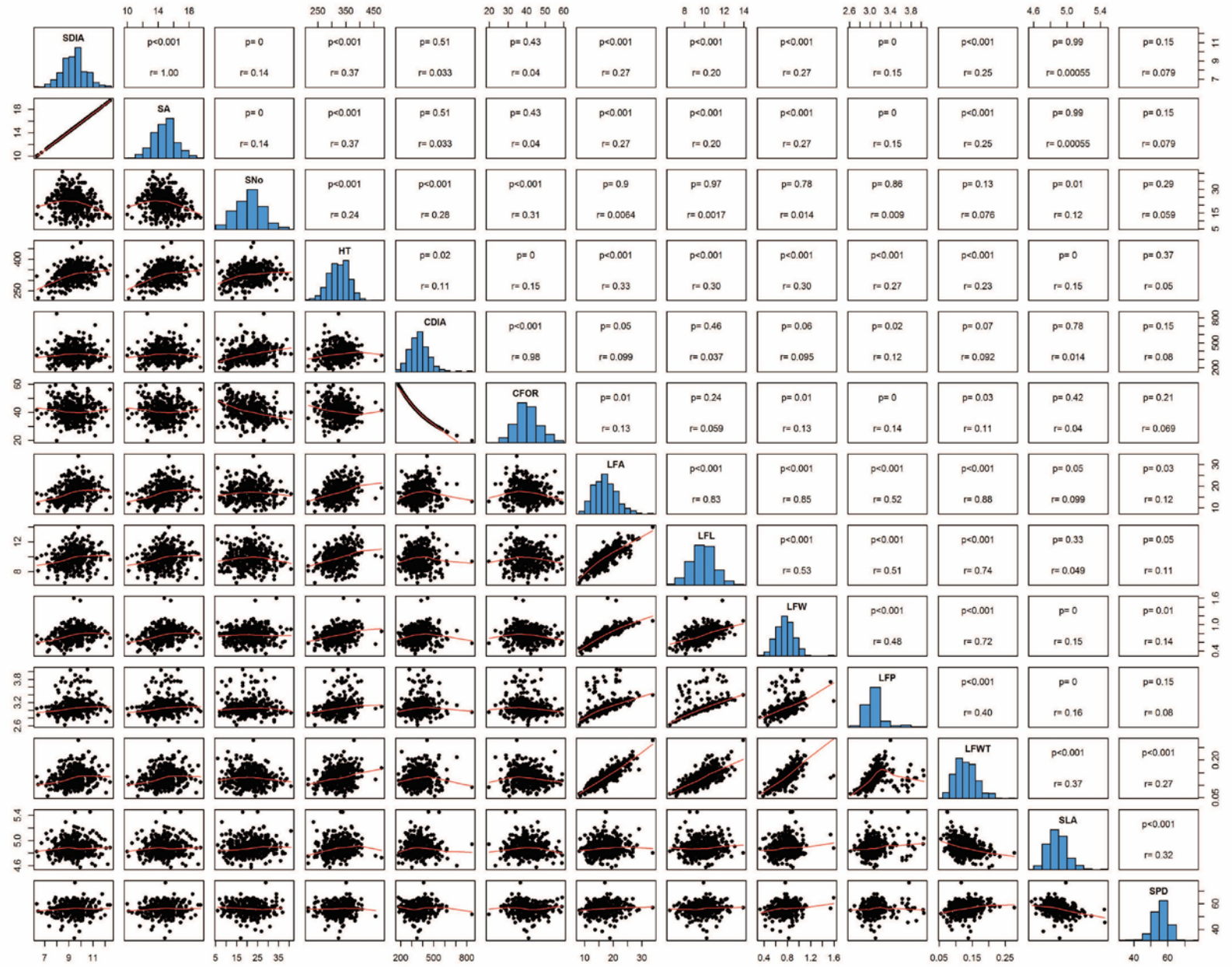
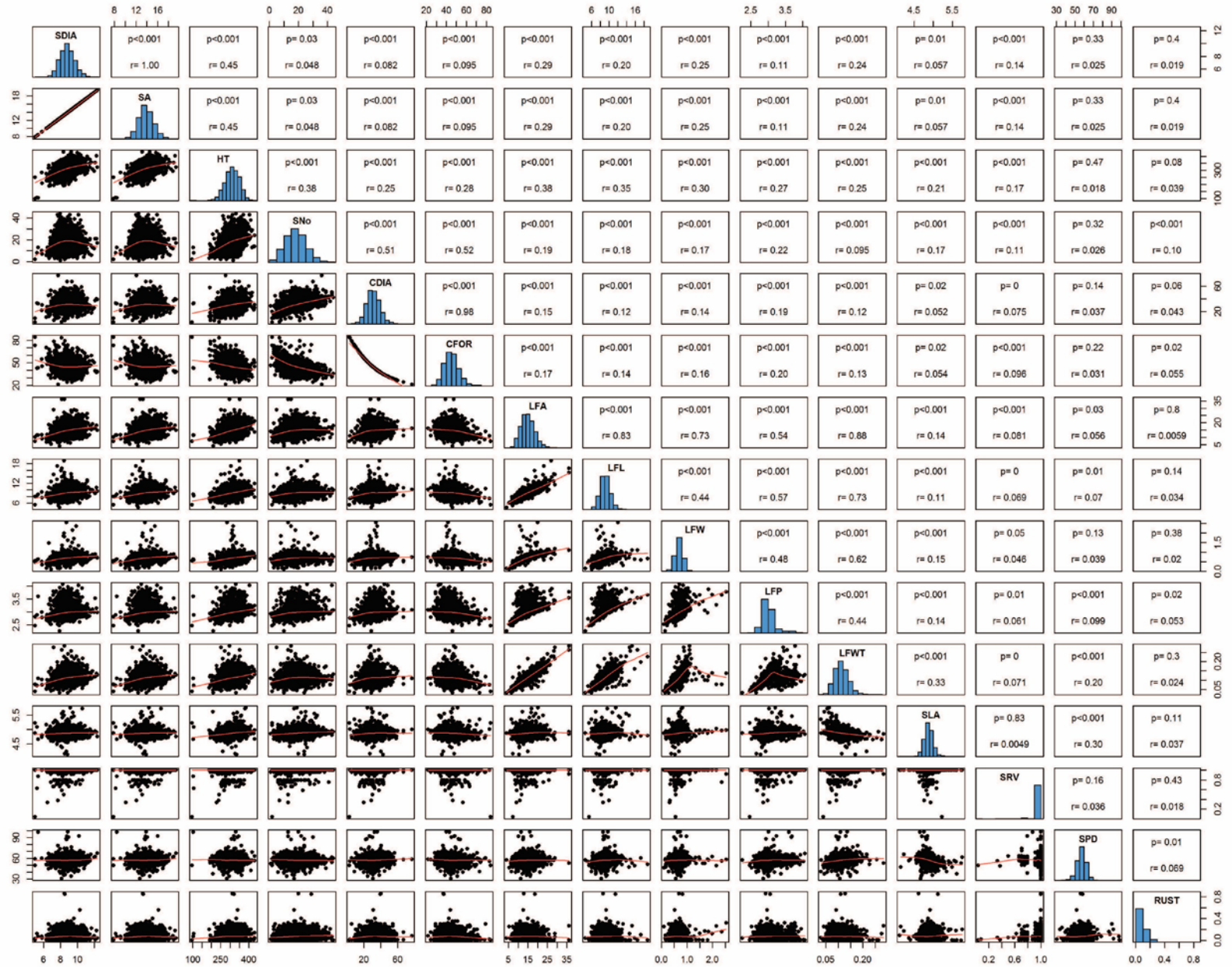


Figure A3.3 Matrix of all pair-wise comparisons between traits measured in 2015 for the *S. purpurea* F₂ family



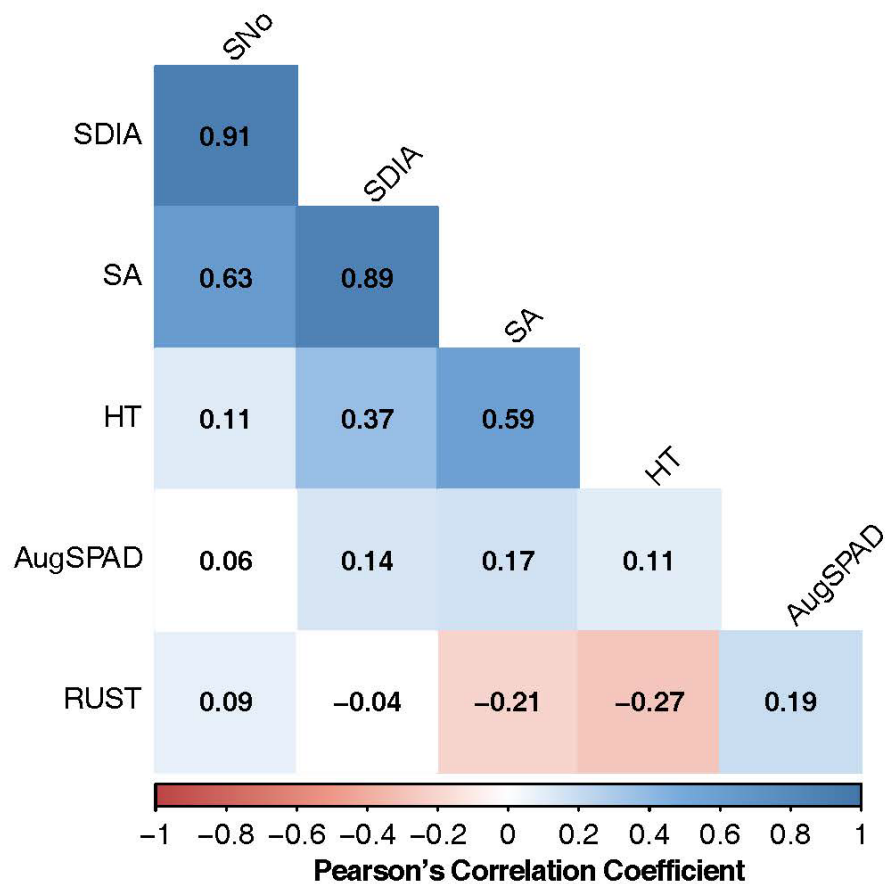


Figure A3.4 Correlation heatmap of traits measured in the diverse collection for the 2015 growing season. Colored boxes indicate significant correlations at $P < 0.05$, where correlation coefficients of -1 are indicated by dark red and 1 shown as dark blue.

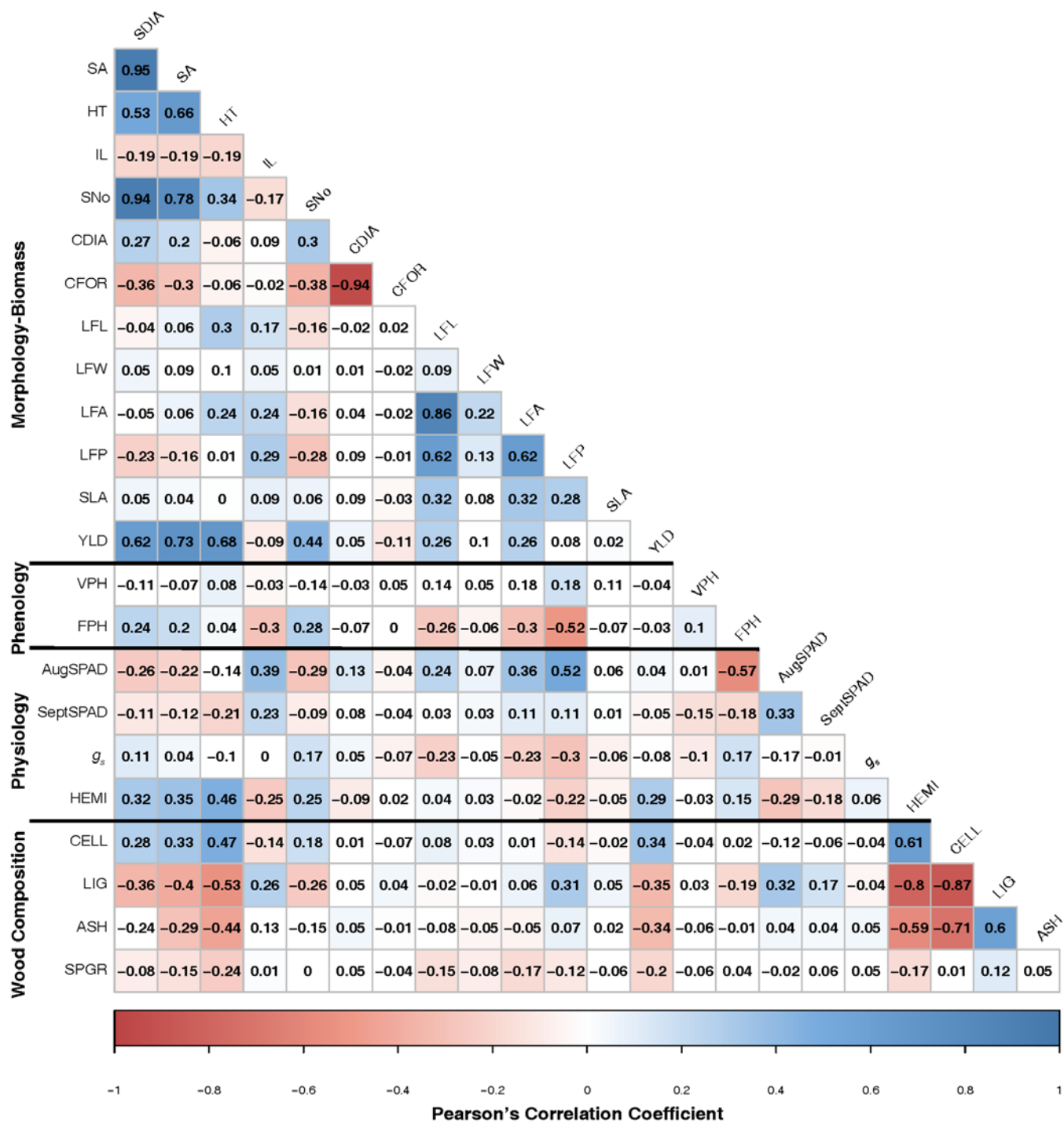


Figure A3.5 Correlation heatmap showing Pearson's correlation coefficients (r) for all *S. purpurea* accessions ($n=110$). Phenotypic means were averaged across years (2013-2014) when appropriate. Traits shown are divided by category as listed in Table 3.1. Colored boxes indicate significant correlations at $P<0.05$, where correlation coefficients of -1 are indicated by dark red and 1 shown as dark blue. All pairwise comparisons between traits by year and location are shown in Figure A3.1.

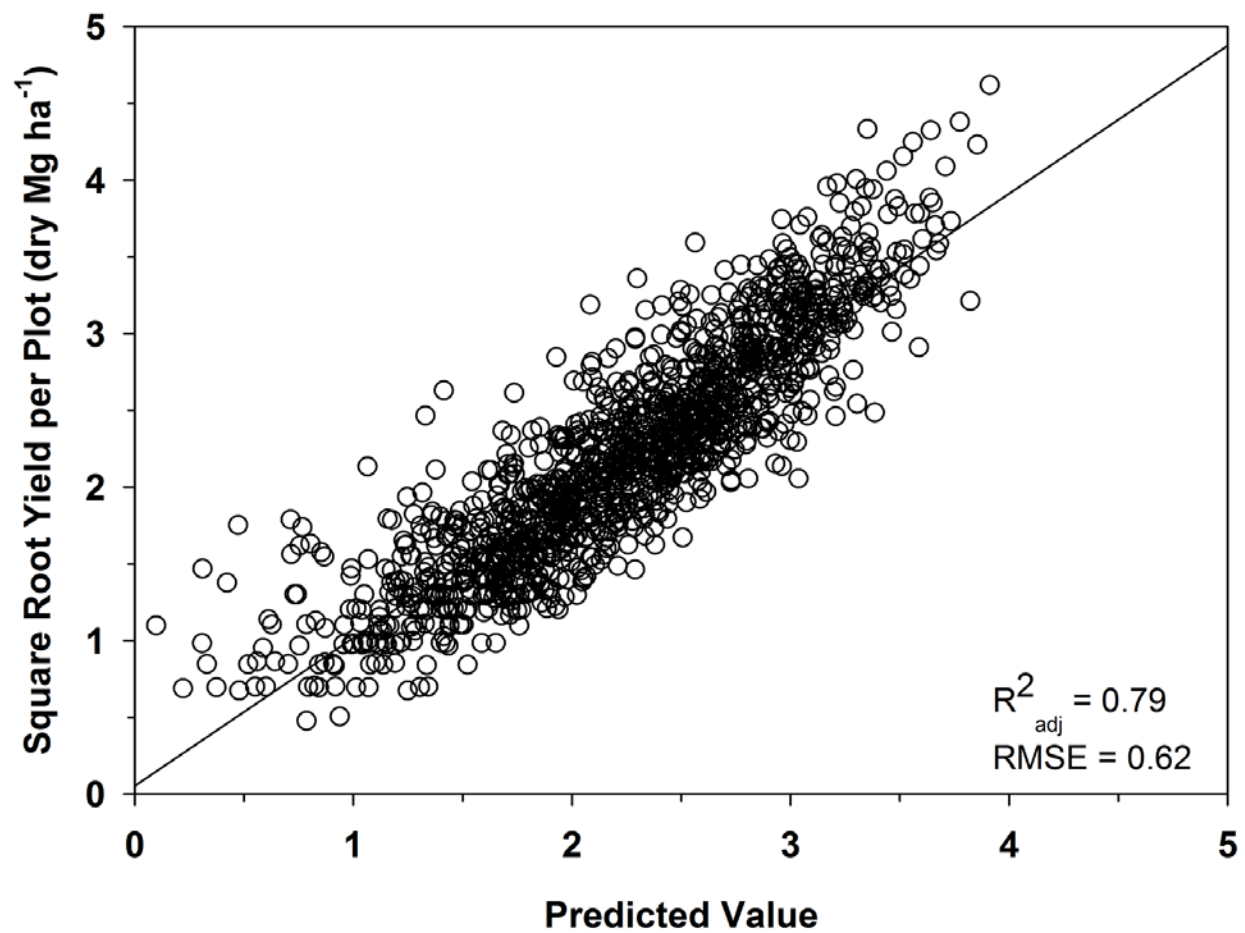


Figure A3.6 Multiple linear regression model for estimating second year post-coppice biomass yield from annual measurements

APPENDIX TO CHAPTER 4

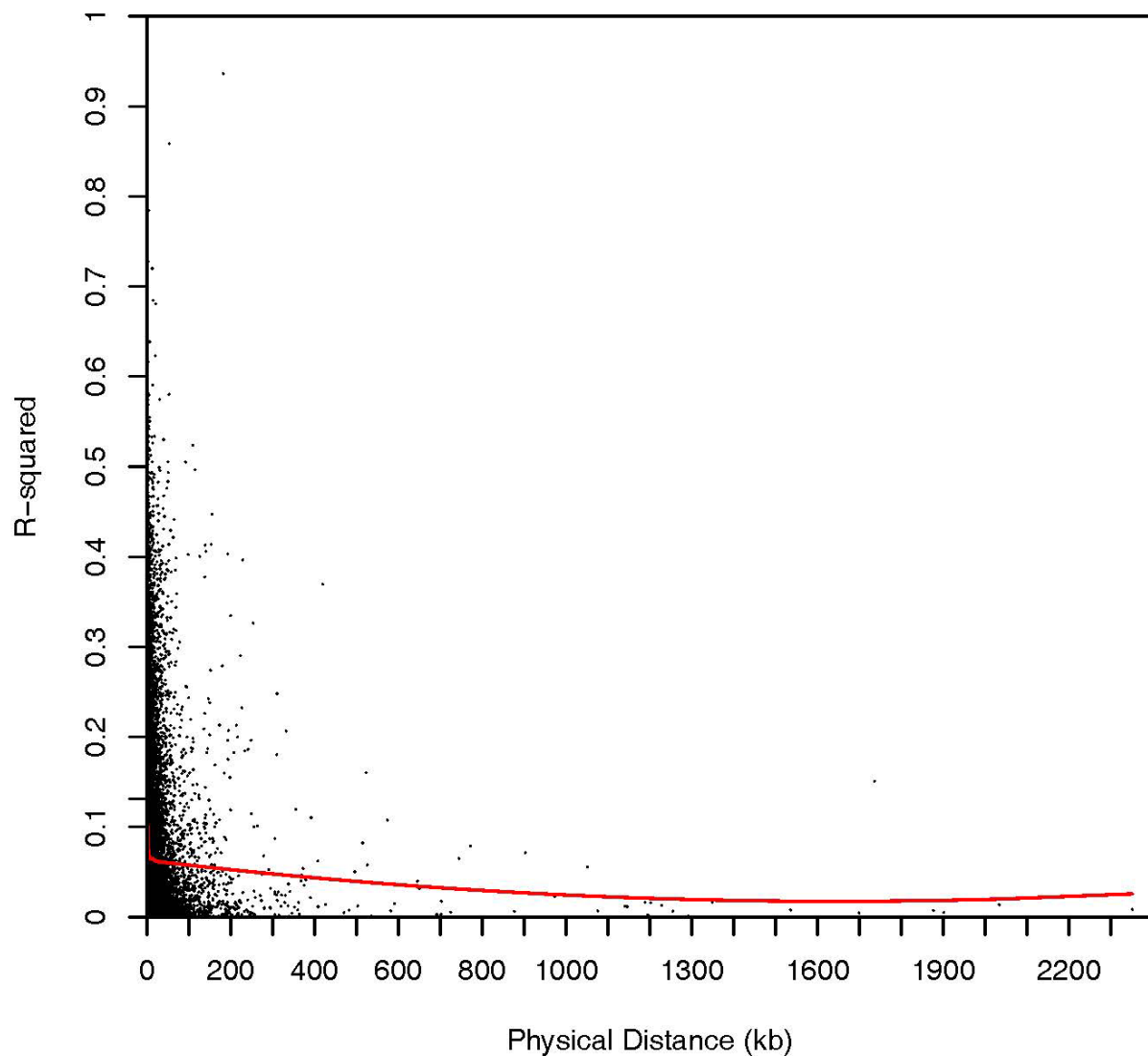


Figure A4.1 Genome-wide linkage disequilibrium (LD) in *S. purpurea* based on common single-nucleotide polymorphisms of minor allele frequency (MAF<0.05). Black circles correspond to average values of R^2 between physical marker locations (kb). The critical LD was defined at a distance of 1.9 kb.

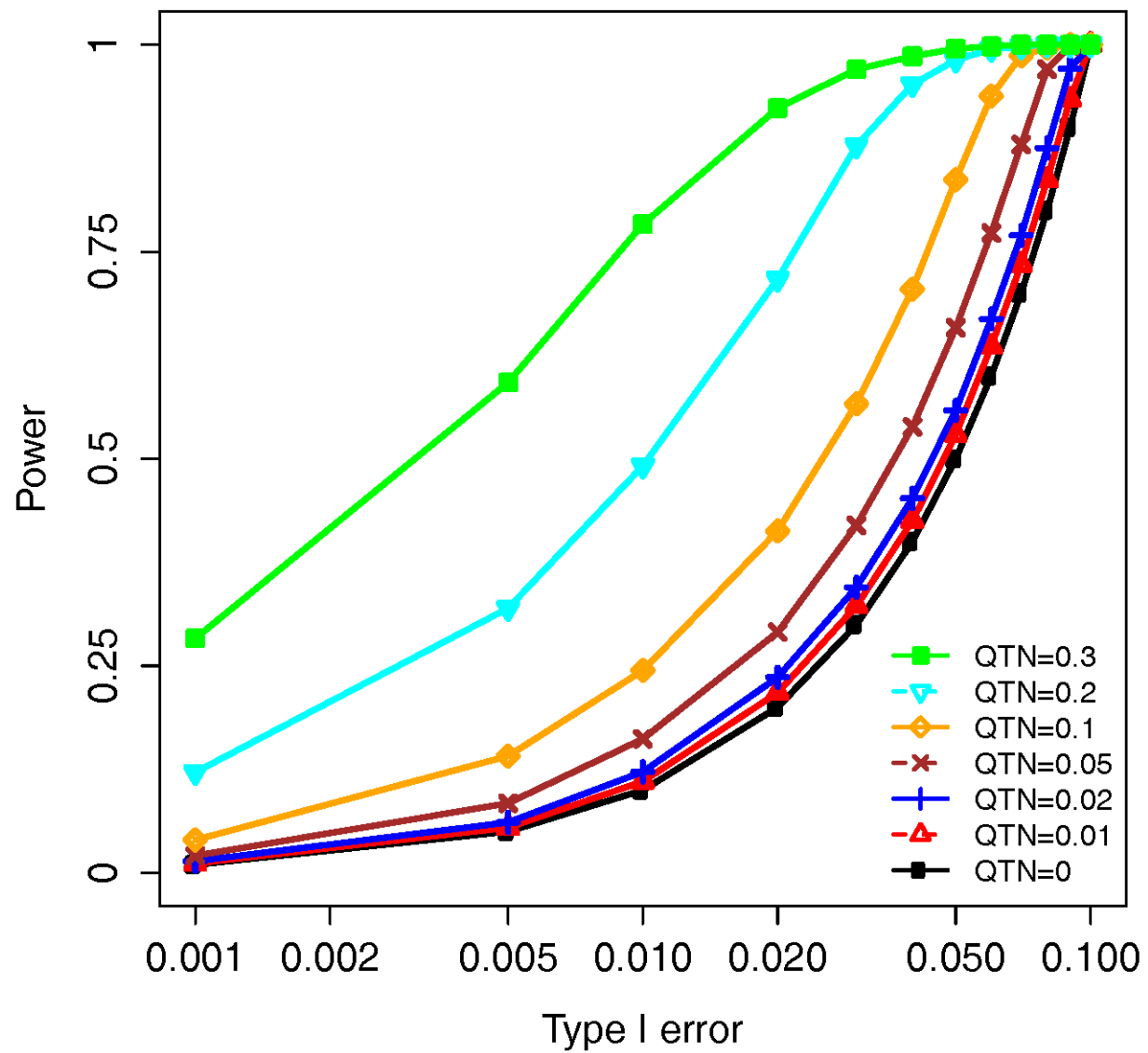


Figure A4.2 Type 1 error plot showing the tradeoff between type I error rate (x -axis) and power (y-axis) for quantitative trait nucleotides (QTNs) with different effect sizes

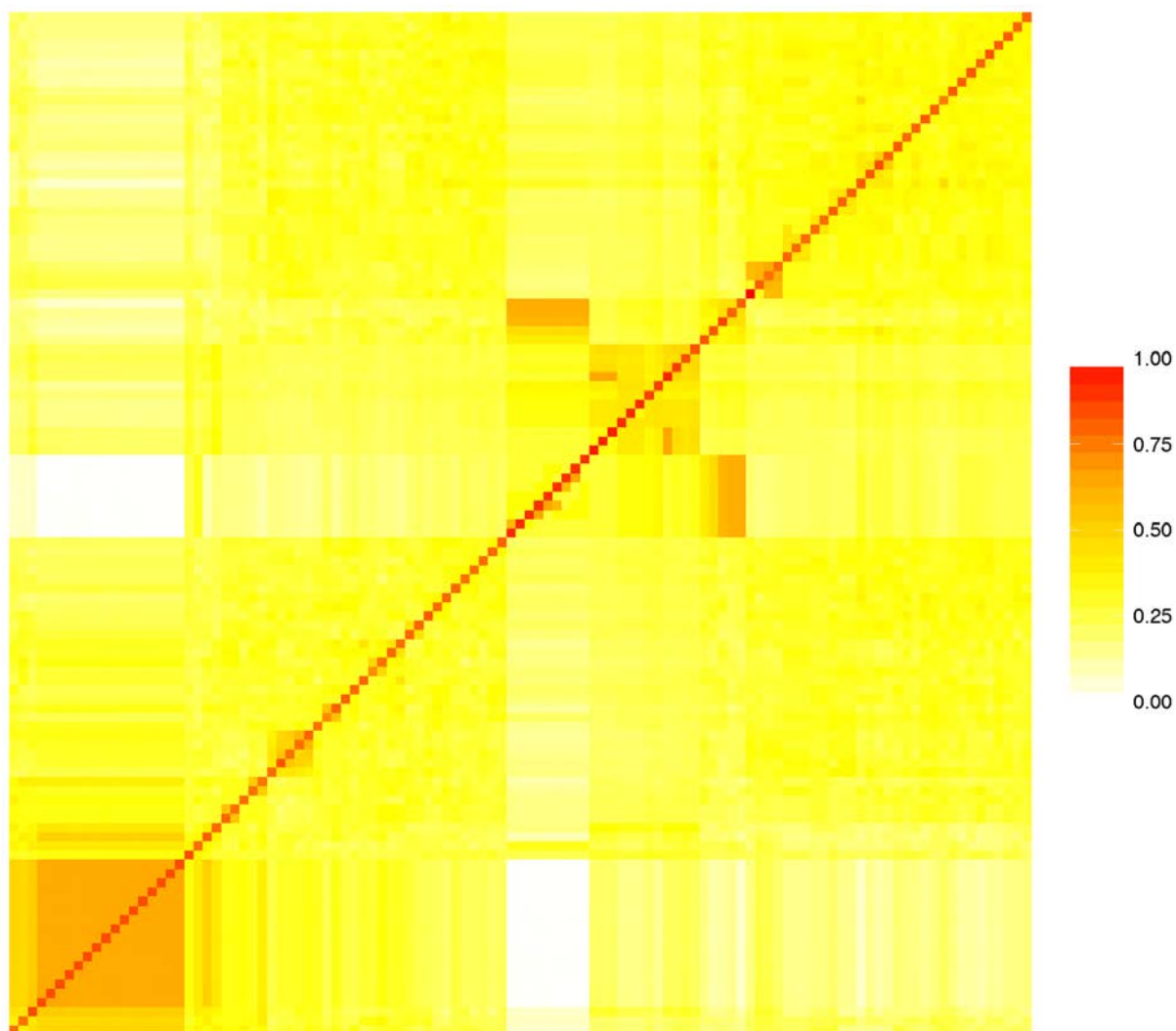


Figure A4.3 Heat map of pairwise kinship among individuals included in the study. Red squares indicate an individual's genetic relatedness to itself and orange blocks indicate genetically similar accession groups by subpopulation

Figure A4.4 Q-Q plots of $-\log_{10}(p\text{-values})$ from mixed model association analyses. Black circles correspond to MLM method, blue circles for CMLM method, and green circles for SUPER method

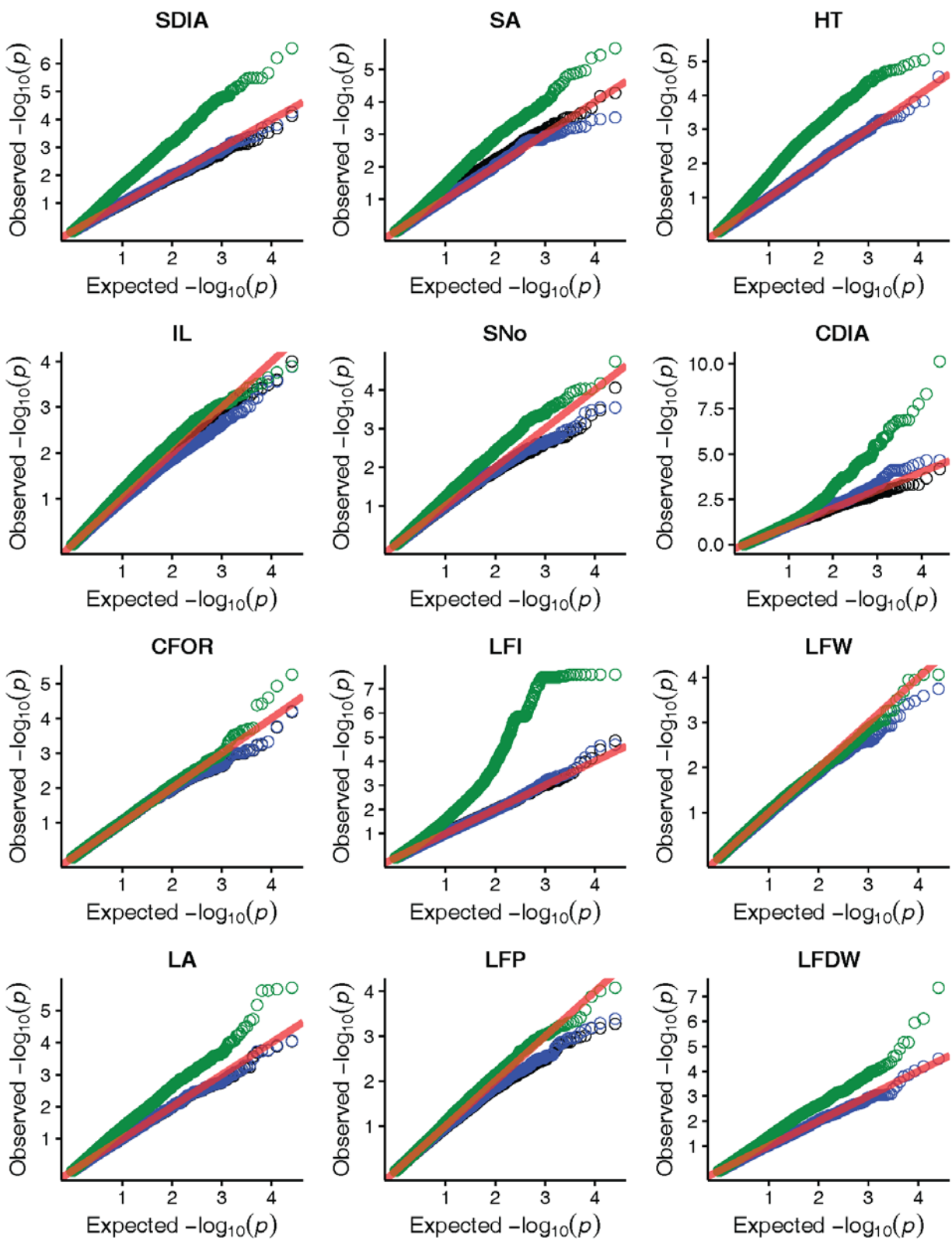


Figure A4.4 (Continued)

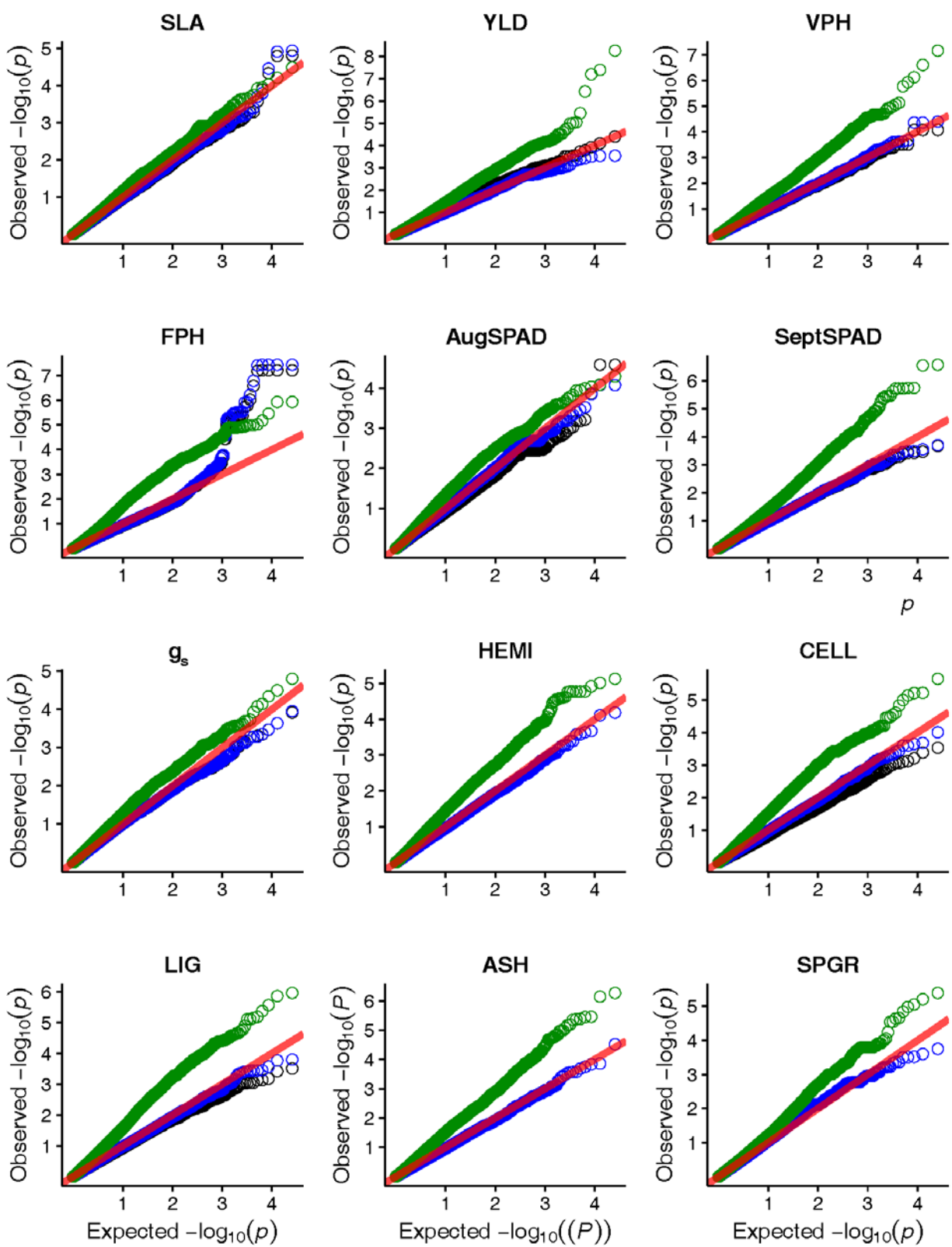
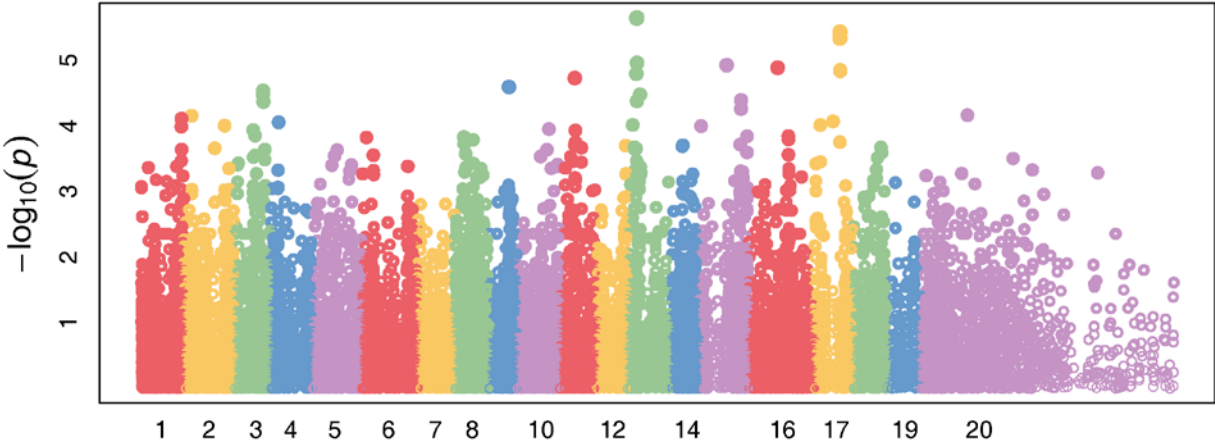
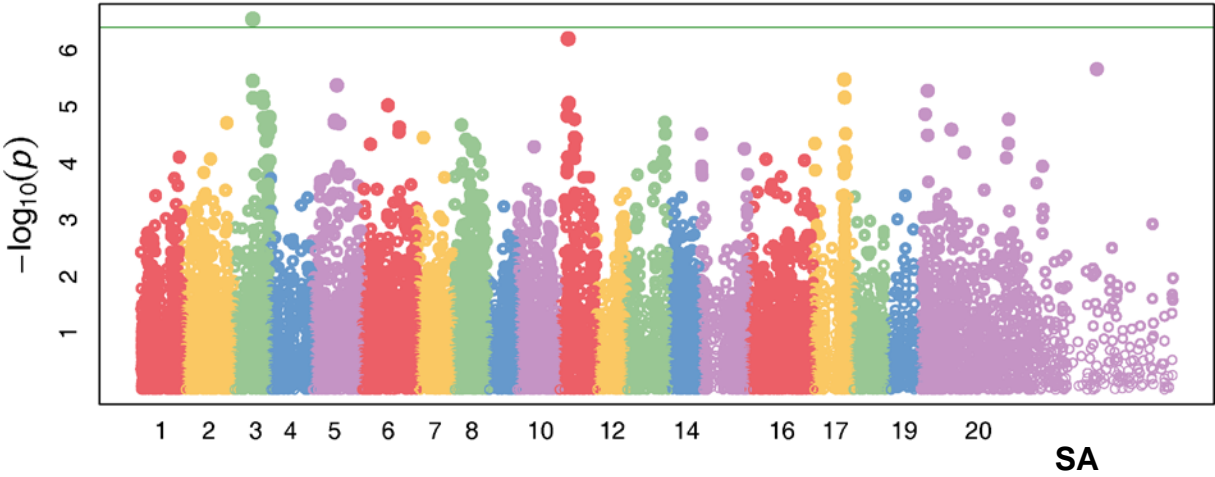


Figure A4.5 Manhattan plots of GWAS results for 24 traits using 25,566 common SNPs throughout the genome. The 20th naïve pseudo-chromosome represents concatenated unassembled scaffolds. Chromosomal locations of $-\log_{10}(p\text{-values})$ for associations at each locus are shown with the green line illustrating the threshold for $\text{FDR} < 0.05$

Figure A4.5 (Continued)

SDIA



HT

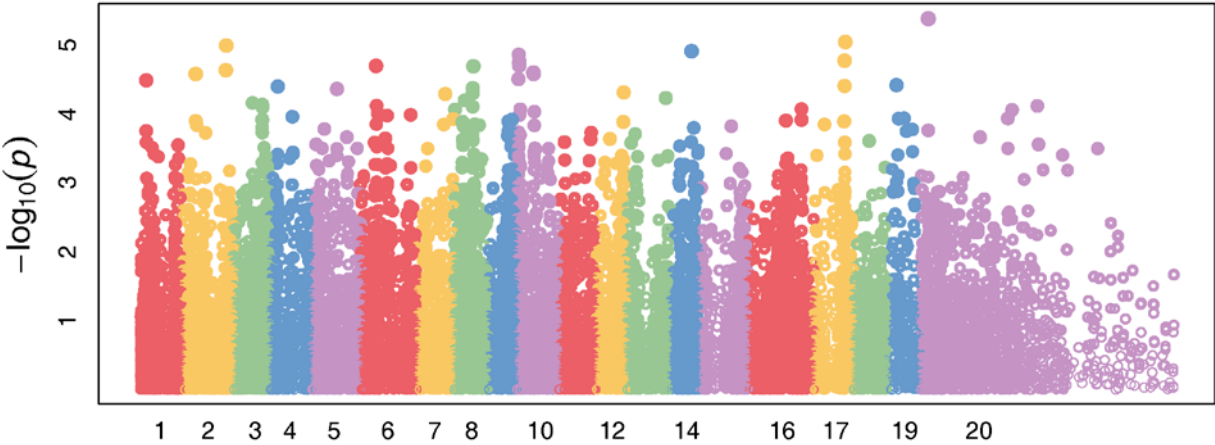
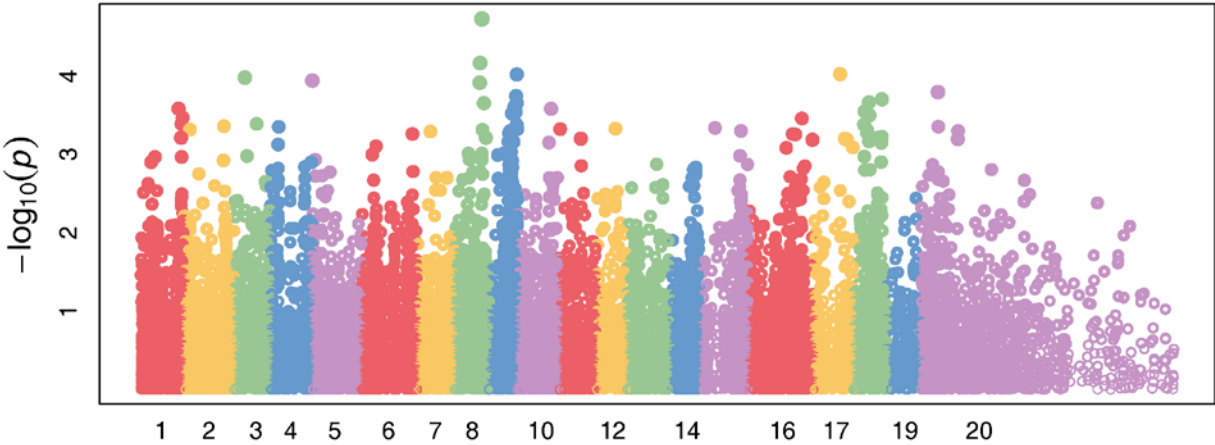
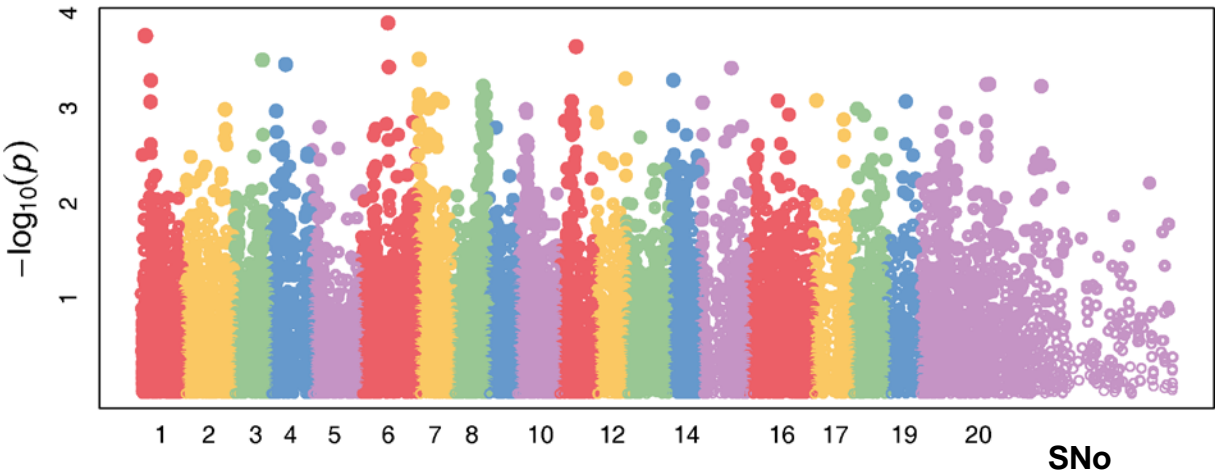


Figure A4.5 (Continued)

IL



CDIA

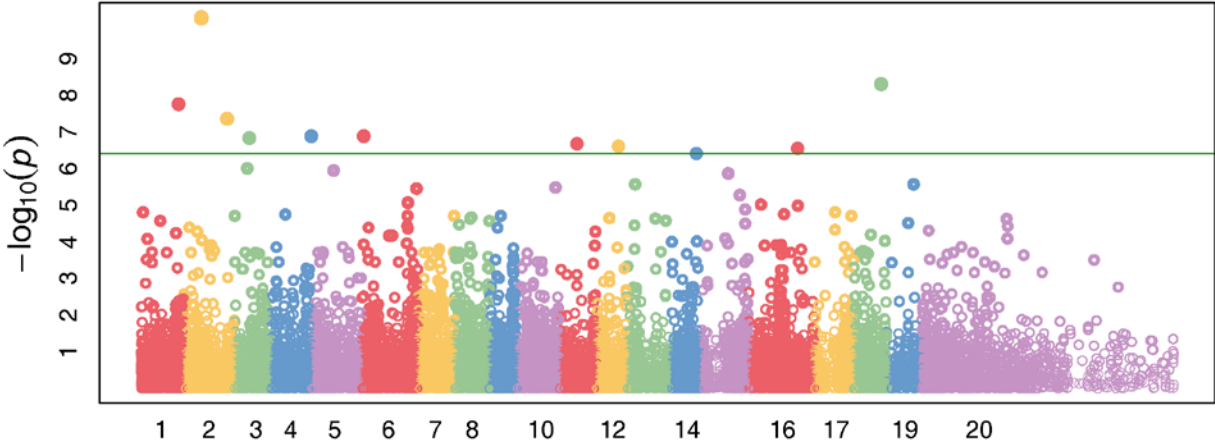
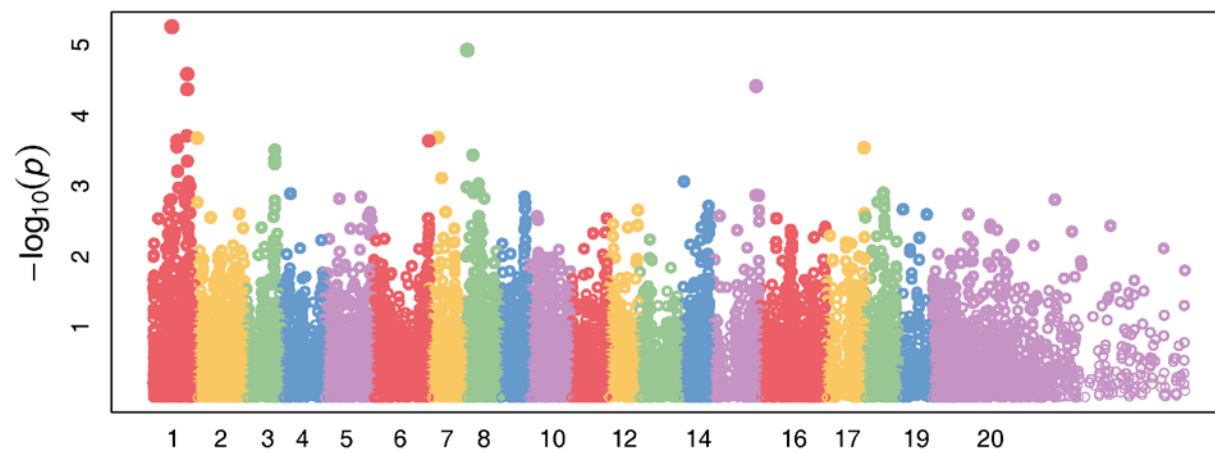
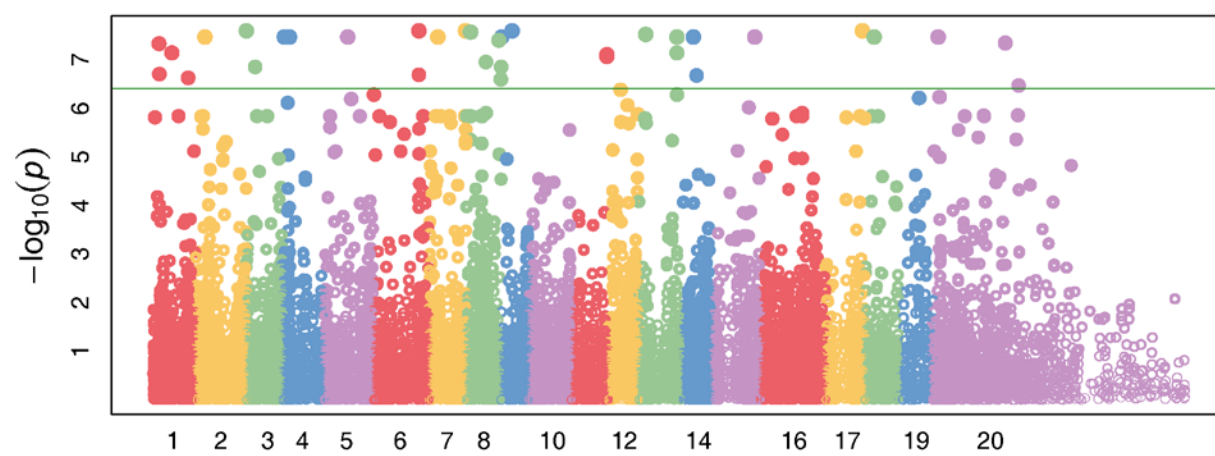


Figure A4.5 (Continued)

CFOR



LFL



LFW

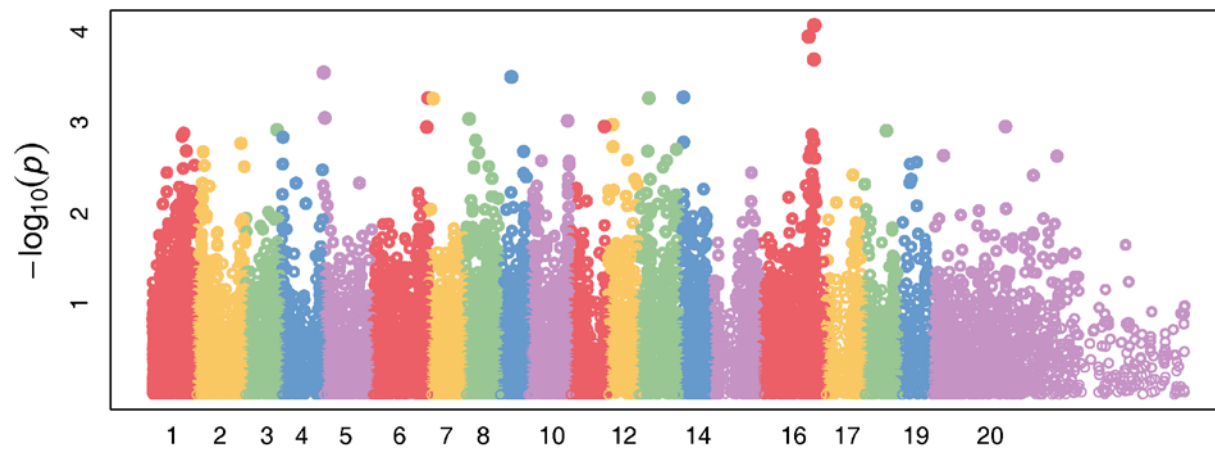
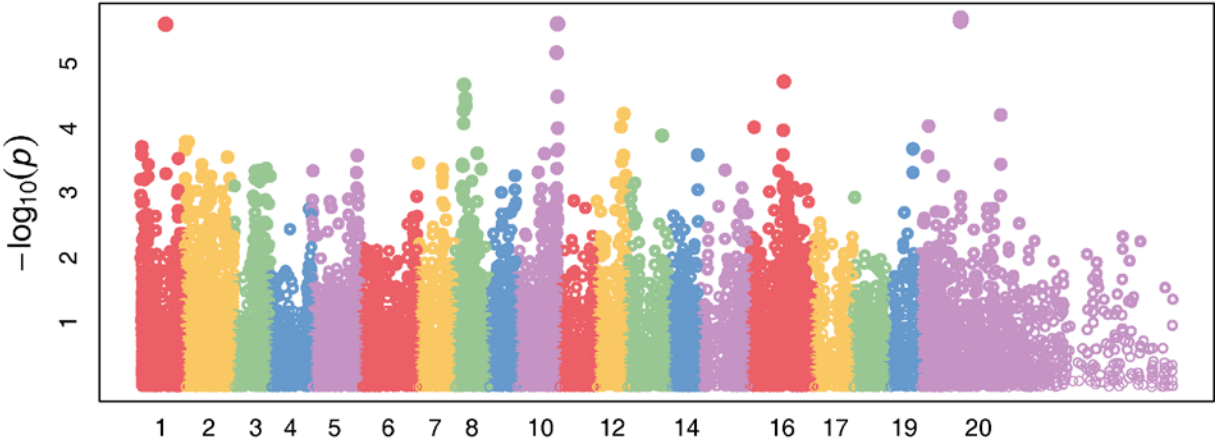
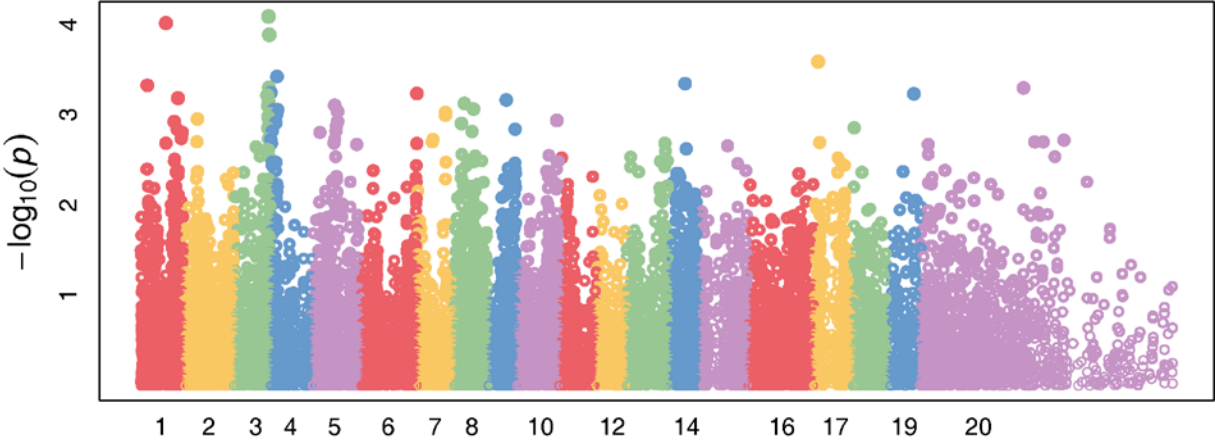


Figure A4.5 (Continued)

LFA



LFP



LDW

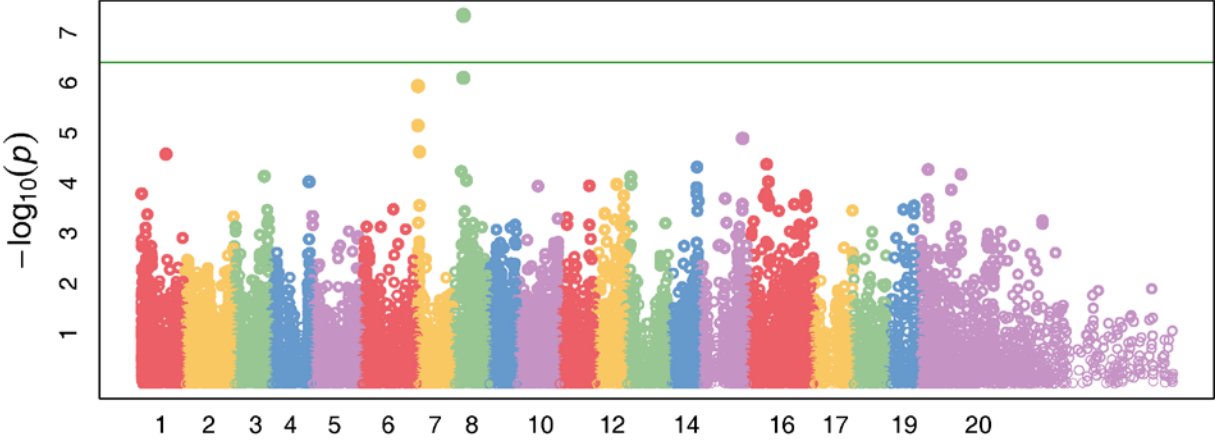
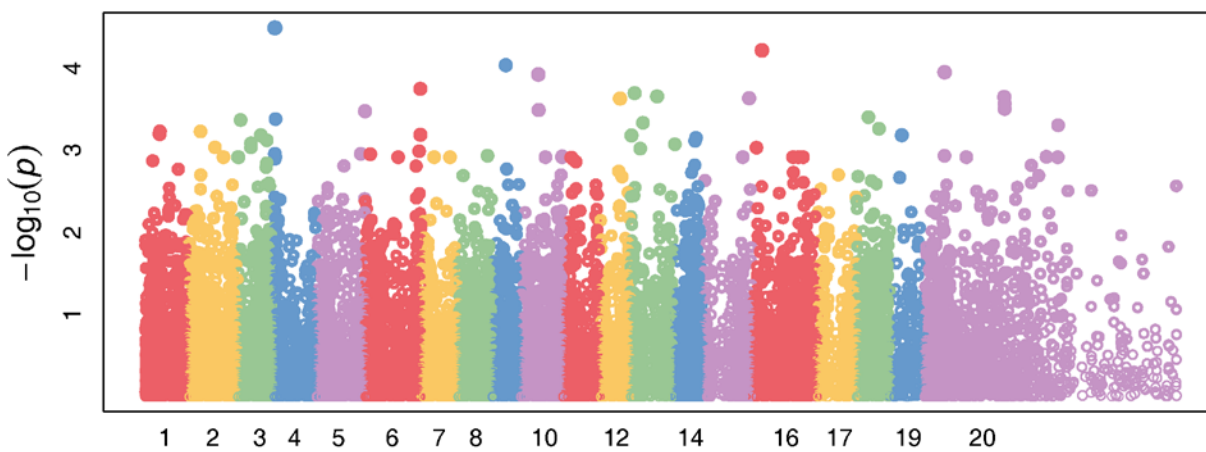
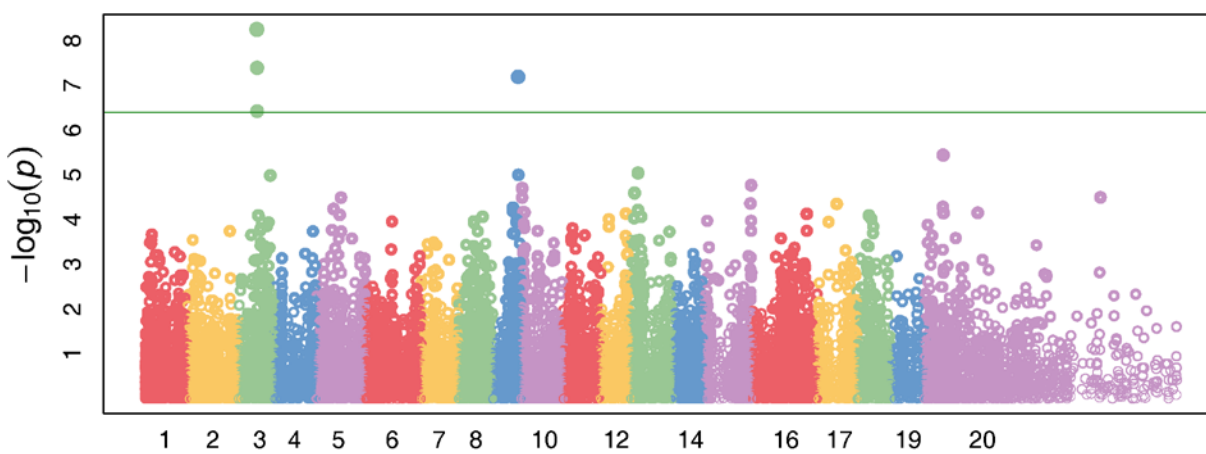


Figure A4.5 (Continued)

SLA



YLD



VPH

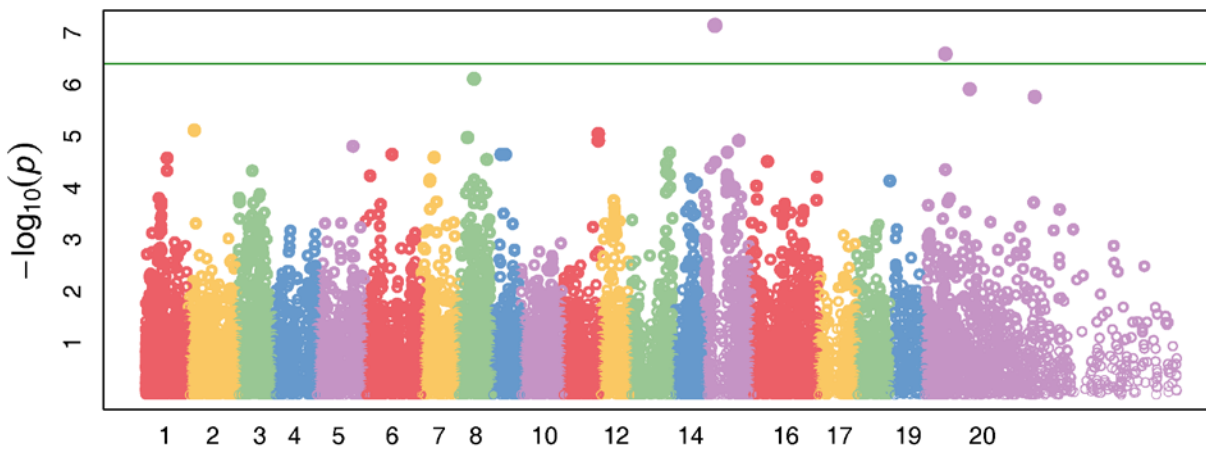
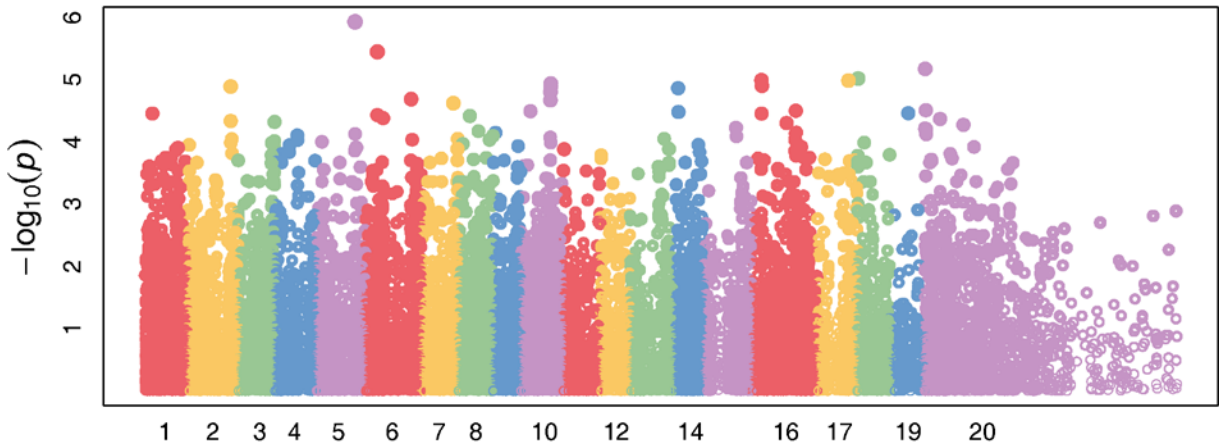
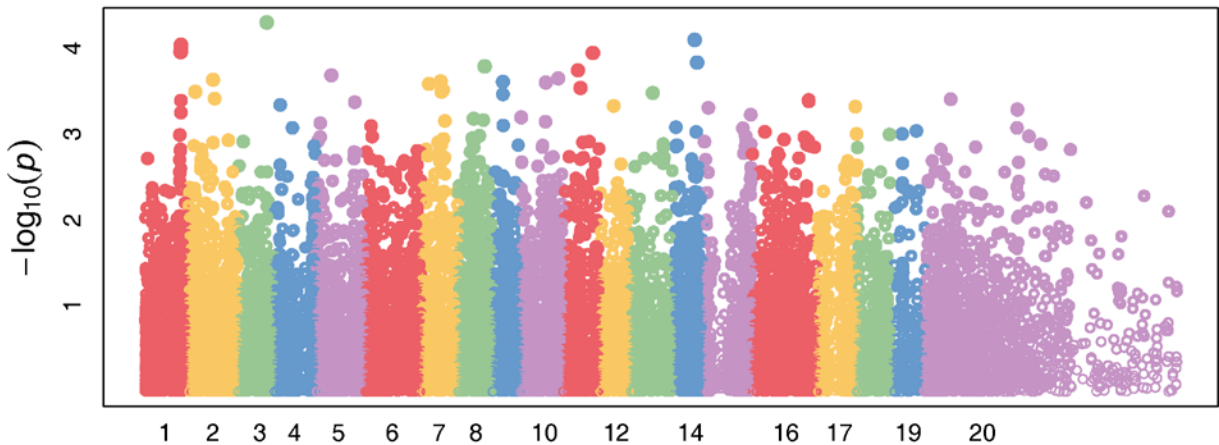


Figure A4.5 (Continued)

FPH



AugSPAD



SeptSPAD

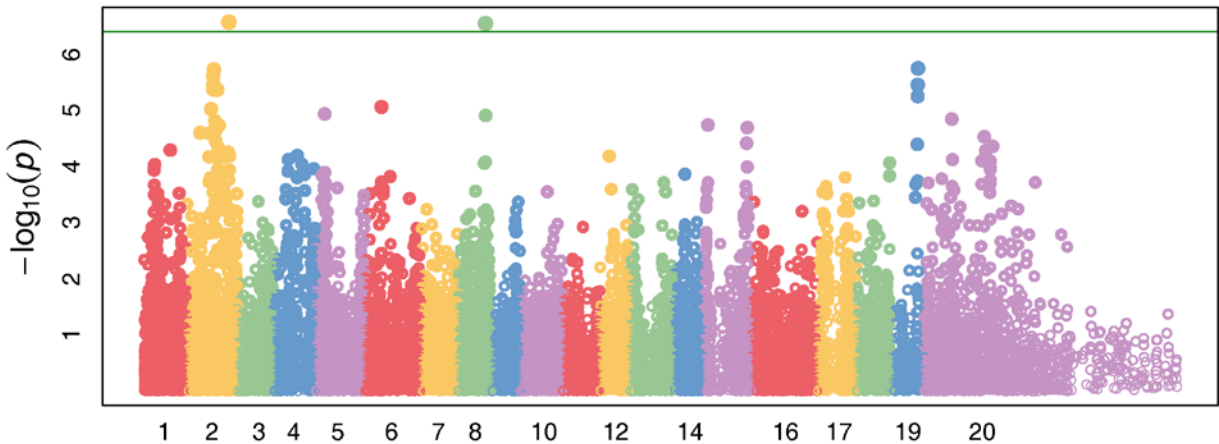
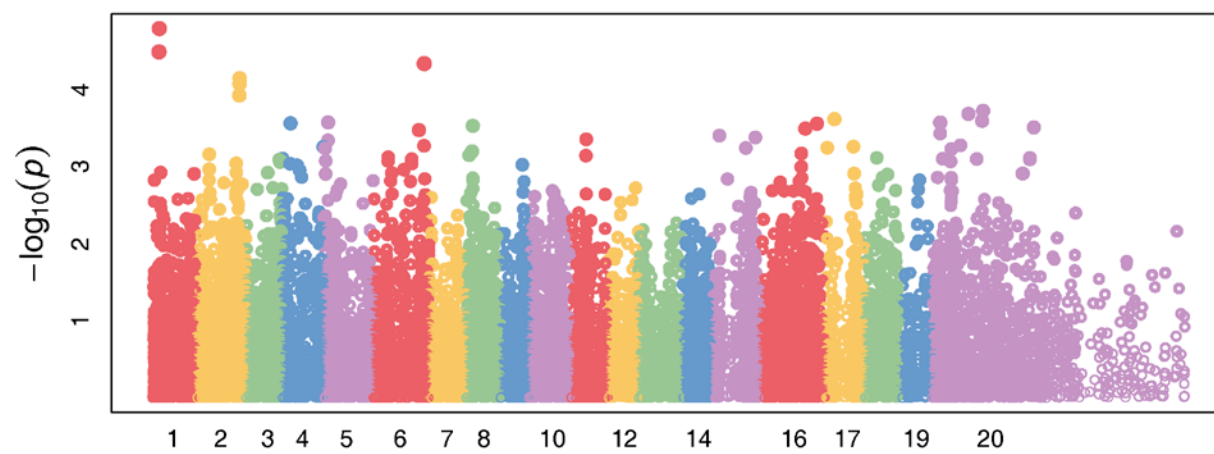
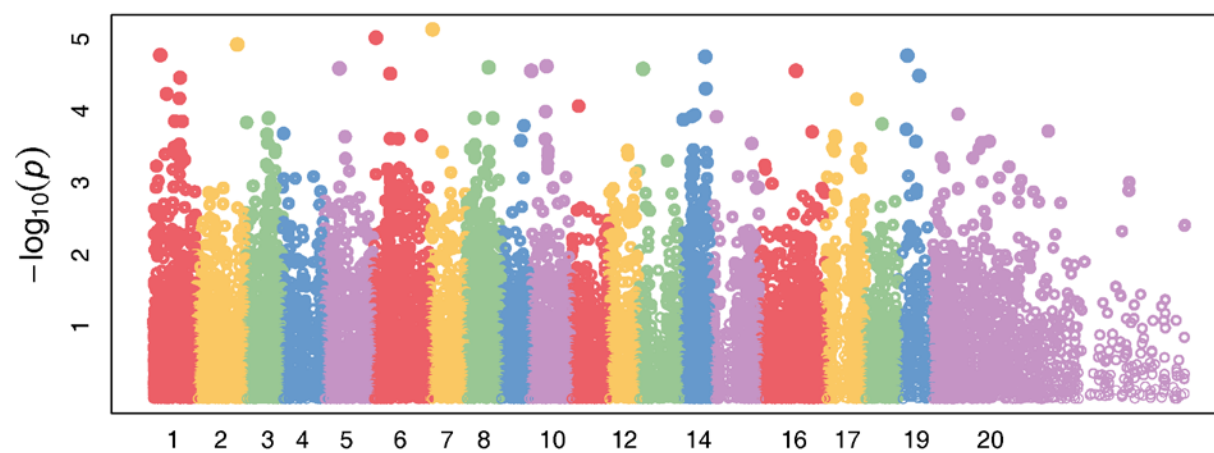


Figure A4.5 (Continued)

g_s



HEMI



CELL

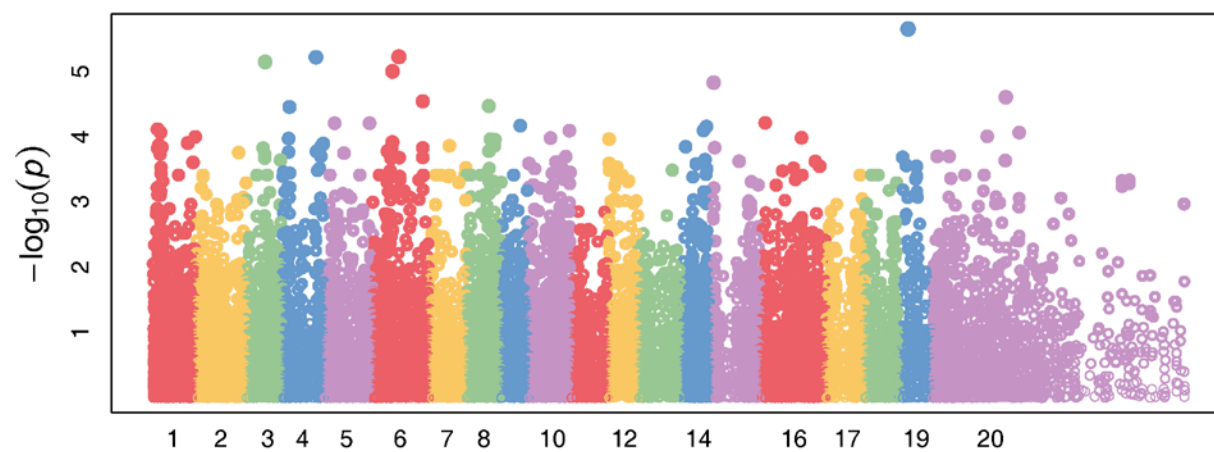
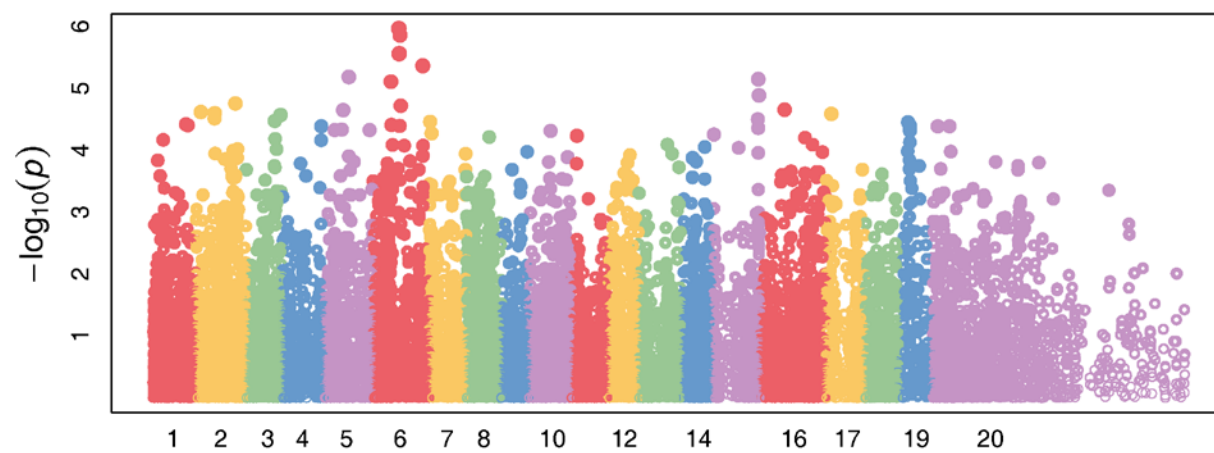
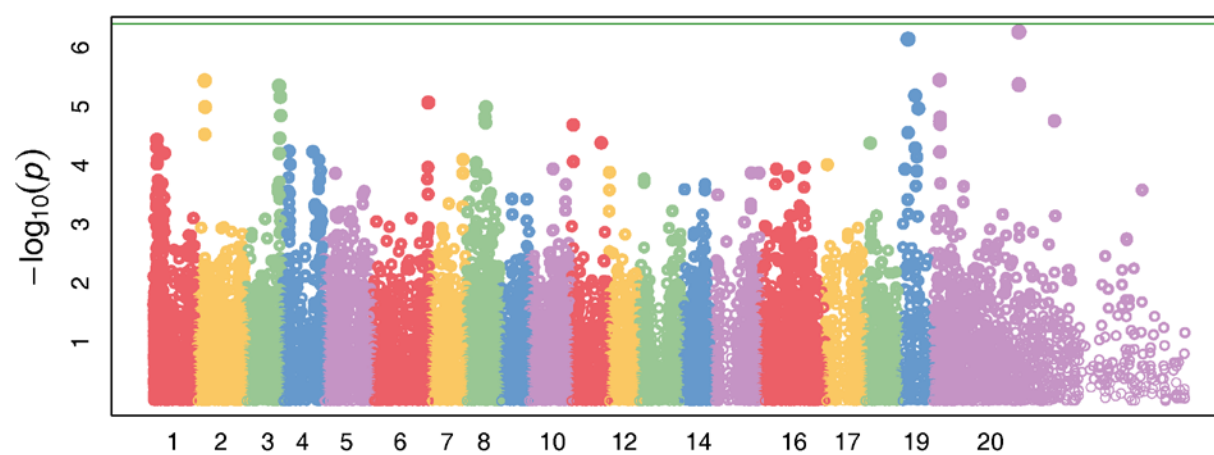


Figure A4.5 (Continued)

LIG



ASH



SPGR

